Sawfly Genomes Reveal Evolutionary Acquisitions That Fostered the Mega-Radiation of Parasitoid and Eusocial Hymenoptera

Oeyen, Jan Philip; Baa-Puyoulet, Patrice; Benoit, Joshua B.; Beukeboom, Leo W.; Bornberg-Bauer, Erich; Buttstedt, Anja; Calevro, Federica; Cash, Elizabeth I.; Chao, Hsu; Charles, Hubert; Chen, Mei-Ju May; Childers, Christopher; Cridge, Andrew G.; Dearden, Peter; Dinh, Huyen; Doddapaneni, Harsha Vardhan; Dolan, Amanda; Donath, Alexander; Dowling, Daniel; Dugan, Shannon; Duncan, Elizabeth; Elpidina, Elena N.; Friedrich, Markus; Geuverink, Elzemiek; Gibson, Joshua D.; Grath, Sonja; Grimmelikhuijzen, Cornelis J. P.; Große-Wilde, Ewald; Gudobba, Cameron; Han, Yi; Hansson, Bill S.; Hauser, Frank; Hughes, Daniel S. T.; Ioannidis, Panagiotis; Jacquin-Joly, Emmanuelle; Jennings, Emily C.; Jones, Jeffery W.; Klasberg, Steffen; Lee, Sandra L.; Lesný, Peter; Lovegrove, Mackenzie; Martin, Sebastian; Martynov, Alexander G.; Mayer, Christoph; Montagné, Nicolas; Moris, Victoria C.; Munoz-Torres, Monica; Murali, Shwetha Canchi; Muzny, Donna M.; Oppert, Brenda; Parisot, Nicolas; Pauli, Thomas; Peters, Ralph S.; Petersen, Malte; Pick, Christian; Persyn, Emma; Podsiadlowski, Lars; Poelchau, Monica F.; Provataris, Panagiotis; Qu, Jiaxin; Reijnders, Maarten J. M. F.; Marcus von Reumont, Björn; Rosendale, Andrew J.; Simao, Felipe A.; Skelly, John; Sotiropoulos, Alexandros G.; Stahl, Aaron L.; Sumitani, Megumi; Szuter, Elise M.; Tidswell, Olivia; Tsitlakidis, Evangelos; Vedder, Lucia; Waterhouse, Robert M.; Werren, John H.; Wilbrandt, Jeanne; Worley, Kim C.; Yamamoto, Daisuke S.; van de Zande, Louis; Zdobnov, Evgeny M.; Ziesmann, Tanja; Gibbs, Richard A.; Richards, Stephen; Hatakeyama, Masatsugu; Misof, Bernhard; Niehuis, Oliver

Published in:
Genome Biology and Evolution

DOI:
10.1093/gbe/evaa106

Publication date:
2020

Document version
Publisher's PDF, also known as Version of record

Sawfly Genomes Reveal Evolutionary Acquisitions That Fostered the Mega-Radiation of Parasitoid and Eusocial Hymenoptera

Jan Philip Oeyen[1,50,*], Patrice Baa-Puyoulet[2], Joshua B. Benoit[3], Leo W. Beukeboom[4], Erich Bornberg-Bauer[5], Anja Buttstedt[6], Federica Calevro[2], Elizabeth I. Cash[7,8], Hsu Chao[9], Hubert Charles[2], Mei-Ju May Chen[10], Christopher Childers[11], Andrew G. Cridge[12], Peter Dearden[12], Huyen Dinh[9], Harsha Vardhan Doddapaneni[9], Amanda Dolan[13], Alexander Donath[1], Daniel Dowling[5], Shannon Dugan[9], Elizabeth Duncan[14], Elena N. Elpidina[15], Markus Friedrich[16], Elzemiek Geuverink[4], Joshua D. Gibson[17,18], Sonja Grath[19], Cornelis J.P. Grimmelikhuijzen[20], Ewald Große-Wilde[21,22], Cameron Gudobba[23], Yi Han[9], Bill S. Hansson[21], Frank Hauser[20], Daniel S.T. Hughes[9], Panagiotis Ioannidis[24,25,26], Emmanuelle Jacquin-Joly[27], Emily C. Jennings[3], Jeffery W. Jones[28], Steffen Klasberg[5], Sandra L. Lee[9], Peter Lesný[29], Mackenzie Lovegrove[12], Sebastian Martin[29], Alexander G. Martynov[30], Christoph Mayer[1], Nicolas Montagné[31], Victoria C. Moris[32], Monica Munoz-Torres[33], Shwetha Canchi Murali[9], Donna M. Muzny[9], Brenda Oppert[34], Nicolas Parisot[2], Thomas Pauli[32], Ralph S. Peters[35], Malte Petersen[1,36], Christian Pick[37], Emma Persyn[31], Lars Podsiadlowski[1], Monica F. Poelchau[11], Panagiotis Provataris[1], Jiaxin Qu[9], Maarten J.M.F. Reijnders[38,39], Björn Marcus von Reumont[40,41], Andrew J. Rosendale[3], Felipe A. Simao[24,25], John Skelly[12], Alexandros G. Sotiropoulos[20], Aaron L. Stahl[3,42], Megumi Sumitani[43], Elise M. Szuter[7], Olivia Tidswell[44,45], Evangelos Tsitlakidis[20], Lucia Vedder[46], Robert M. Waterhouse[38,39], John H. Werren[13], Jeanne Wilbrandt[1,47], Kim C. Worley[9], Daisuke S. Yamamoto[48], Louis van de Zande[4], Evgeny M. Zdobnov[24,25], Tanja Ziesmann[1], Richard A. Gibbs[9], Stephen Richards[9], Masatsugu Hatakeyama[49], Bernhard Misof[1,*], and Oliver Niehuis[32,*]

[1]Center for Molecular Biodiversity Research, Zoologisches Forschungsmuseum Alexander Koenig, Bonn, Germany

[2]INSA-Lyon, INRAE, BF2I, UMR0203, Université de Lyon, Villeurbanne, France

[3]Department of Biological Sciences, University of Cincinnati

[4]Groningen Institute for Evolutionary Life Sciences, University of Groningen, The Netherlands

[5]Institute for Evolution and Biodiversity, University of Münster, Germany

[6]B CUBE—Center for Molecular Bioengineering, Technische Universität Dresden, Germany

[7]School of Life Sciences, College of Liberal Arts and Sciences, Arizona State University

[8]Department of Environmental Science, Policy, and Management, College of Natural Resources, University of California, Berkeley

[9]Human Genome Sequencing Center, Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, Texas

[10]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan

[11]USDA-ARS, National Agricultural Library, Beltsville, Maryland

[12]Genomics Aotearoa and Biochemistry Department, University of Otago, Dunedin, New Zealand

[13]Department of Biology, University of Rochester

[14]School of Biology, Faculty of Biological Sciences, University of Leeds, United Kingdom

[15]A.N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Russia

[16]Department of Biological Sciences, Wayne State University, Detroit

[17]Department of Biology, Georgia Southern University, Statesboro

[18]Department of Entomology, Purdue University, West Lafayette

[19]Division of Evolutionary Biology, Faculty of Biology, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany

[20]Department of Biology, University of Copenhagen, Denmark
[21]Department of Evolutionary Neuroethology, Max-Planck-Institute for Chemical Ecology, Jena, Germany
[22]Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague (CULS), Praha 6—Suchdol, Czech Republic
[23]Department of Psychiatry and Behavioral Neuroscience, University of Chicago
[24]Department of Genetic Medicine and Development, University of Geneva Medical School, Switzerland
[25]Swiss Institute of Bioinformatics, Geneva, Switzerland
[26]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology—Hellas, Heraklion, Crete, Greece
[27]INRAE, CNRS, IRD, UPEC, Univ. P7, Institute of Ecology and Environmental Sciences of Paris, Sorbonne Université, Versailles, France
[28]Department of Biological Sciences, Oakland University, Rochester
[29]Institute of Evolutionary Biology and Ecology, Zoology and Evolutionary Biology, University of Bonn, Germany
[30]Center of Life Sciences, Skolkovo Institute of Science and Technology, Russia
[31]INRAE, CNRS, IRD, UPEC, Univ. P7, Institute of Ecology and Environmental Sciences of Paris, Sorbonne Université, Paris, France
[32]Department of Evolutionary Biology and Ecology, Institute of Biology I (Zoology), Albert Ludwig University Freiburg, Germany
[33]Berkeley Bioinformatics Open-source Projects (BBOP), Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California
[34]USDA Agricultural Research Service, Center for Grain and Animal Health Research, Manhattan, Kansas
[35]Arthropoda Department, Center for Taxonomy and Evolutionary Research, Zoologisches Forschungsmuseum Alexander Koenig, Bonn, Germany
[36]Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany
[37]Zoological Institute, University of Hamburg, Germany
[38]Department of Ecology and Evolution, University of Lausanne, Switzerland
[39]Swiss Institute of Bioinformatics, Lausanne, Switzerland
[40]Institute for Insect Biotechnology, University of Gießen, Germany
[41]Center for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt, Germany
[42]Department of Neuroscience, The Scripps Research Institute, Jupiter, Florida
[43]Transgenic Silkworm Research Unit, Division of Biotechnology, Institute of Agrobiological Sciences, National Agriculture and Food Research Organization (NARO), Owashi, Tsukuba, Japan
[44]Biochemistry Department, University of Otago, Dunedin, New Zealand
[45]Zoology Department, University of Cambridge, United Kingdom
[46]Center for Bioinformatics Tübingen (ZBIT), University of Tübingen, Germany
[47]Computational Biology Group, Leibniz Institute on Aging—Fritz Lipmann Institute, Jena, Germany
[48]Division of Medical Zoology, Department of Infection and Immunity, Jichi Medical University, Yakushiji, Shimotsuke, Japan
[49]Insect Genome Research and Engineering Unit, Division of Applied Genetics, Institute of Agrobiological Sciences, NARO, Owashi, Tsukuba, Japan
[50]Lead Contact

*Corresponding authors: E-mails: janphilipoyen@gmail.com; b.misof@leibniz-zfmk.de; oliver.niehuis@biologie.uni-freiburg.de.

Accepted: 19 May 2020

Data deposition: The genome sequence assemblies of *Athalia rosae* and *Orussus abietinus* have been deposited at Ag Data Commons archives (doi:10.15482/USDA.ADC/1459563 and 10.15482/USDA.ADC/1459569, respectively). The raw genome sequencing reads have been deposited at NCBI Sequence Read Archives (SRA) under the accessions SRX276605–9 and SRX330912–15, respectively. The raw whole-body transcriptome sequencing reads of *At. rosae* and *O. abietinus* have been deposited at SRA under the accessions SRX903108–9 and SRX906143–4, respectively. The raw antennal transcriptome sequencing reads of *At. rosae* and *O. abietinus* have been deposited at SRA under the accessions ERX1404801–2 and ERX1404803, respectively. The automated genome annotation of the *At. rosae* and *O. abietinus* draft genomes have been deposited at Ag Data Commons archives (doi:10.15482/USDA.ADC/1459565 and 10.15482/USDA.ADC/1459561, respectively). The official gene sets (merged automated and manual annotations) of *At. rosae* and *O. abietinus* have been deposited at Ag Data Commons archives (doi:10.15482/USDA.ADC/1459566 and 10.15482/USDA.ADC/1459558, respectively). The metabolism databases of *At. rosae*, *O. abietinus*, and *N. vitripennis* have been included in the ArthropodaCyc database collection (Baa-Puyoulet P, et al. 2016. ArthropodaCyc: a CycADS powered collection of BioCyc databases to analyse and compare metabolism of arthropods. Database 2016:baw081) and are available under http://arthropodacyc.cycadsys.org/ATHRO/, http://arthropodacyc.cycadsys.org/ORUAB/, and http://arthropodacyc.cycadsys.org/NASVI/.

Abstract

The tremendous diversity of Hymenoptera is commonly attributed to the evolution of parasitoidism in the last common ancestor of parasitoid sawflies (Orussidae) and wasp-waisted Hymenoptera (Apocrita). However, Apocrita and Orussidae differ dramatically in their species richness, indicating that the diversification of Apocrita was promoted by additional traits. These traits have remained elusive due to a paucity of sawfly genome sequences, in particular those of parasitoid sawflies. Here, we present comparative analyses of draft genomes of the primarily phytophagous sawfly *Athalia rosae* and the parasitoid sawfly *Orussus abietinus*. Our analyses revealed that the ancestral hymenopteran genome exhibited traits that were previously considered unique to eusocial Apocrita (e.g., low transposable element content and activity) and a wider gene repertoire than previously thought (e.g., genes for CO_2 detection). Moreover, we discovered that Apocrita evolved a significantly larger array of odorant receptors than sawflies, which could be relevant to the remarkable diversification of Apocrita by enabling efficient detection and reliable identification of hosts.

Key words: hexamerin, major royal jelly protein, microsynteny, odorant receptor, opsin, phytophagy.

Introduction

Hymenoptera (sawflies, wasps, ants, and bees) represent one of the four mega-diverse insect orders. It is estimated to comprise over one million species and currently includes over 153,000 described species (Aguiar et al. 2013). The transition from an ancestral ectophytophagous lifestyle, retained by the majority of sawflies ("Symphyta"), to parasitoidism, a lifestyle in which a larva develops by feeding upon and killing a single host specimen, is generally considered the most important factor that promoted the diversification of Hymenoptera (Mrinalini and Werren 2017; Peters et al. 2017). Results from phylogenetic analyses imply that this transition occurred only once during the evolution of Hymenoptera: in the stem lineage of the parasitoid sawfly family Orussidae and the wasp-waisted Hymenoptera (Apocrita) (Peters et al. 2017). The transition to a parasitoid lifestyle was associated with the evolution of numerous adaptations in behavior, morphology, and physiology (Whitfield 1998). For example, parasitoids critically depend on their ability to locate hosts, to successfully lay eggs on or in their hosts, to inject venom to immobilize their host and/or to antagonize their hosts' immune response, and to metabolize a nitrogen-rich animal-based diet (as compared with a nitrogen-poor plant-based diet). Intriguingly, however, wasp-waisted Hymenoptera diversified far more (144,593 described species, $>$ 90% of the extant species of Hymenoptera) than parasitoid sawflies (82 described species), indicating that the diversification of the Apocrita was likely promoted by the evolutionary acquisition of traits that parasitoid sawflies lack. Yet, the transition from phytophagy to parasitoidism and the factors contributing to the massive speciation of Apocrita have remained largely unstudied. The tremendous diversity, as well as the ecological and economical importance of Hymenoptera, have led the order to be the focus of a wealth of taxonomic, evolutionary, and ecological research (Quicke 1997; Grimaldi and Engel 2005; Sharkey 2007; Peters et al. 2017). However, most of the comparative genomic research on Hymenoptera has been focused on Apocrita and especially on the multiple origins of eusociality within this clade. As a result, all but one of the published draft genomes of Hymenoptera refer to species of Apocrita (Branstetter et al. 2018). The only published draft genome of a sawfly is that of the wheat stem sawfly, *Cephus cinctus* (Cephoidea) (Robertson et al. 2018). The larvae of Cephoidea are endophytophagous, feeding on a wide range of large-stemmed grasses, including economically important crops, and show an opportunistic cannibalistic behavior (Beres et al. 2011). As the sister group to Orussidae + Apocrita (Peters et al. 2017), the superfamily Cephoidea represents an important lineage in the hymenopteran tree of life for understanding the possible onsets of parasitoidism. At the same time, the derived ecology of Cephoidea, whose larvae are neither strictly phytophagous nor parasitoid, and its specific systematic position prevent the drawing of major conclusions on the composition of the ancestral genome of (phytophagous) Hymenoptera or on factors contributing to the disparate diversification of the parasitoid Orussidae and Apocrita.

Knowledge of the composition of the ancestral genome of Hymenoptera is fundamental for tracing the evolution of traits within Hymenoptera. In addition, due to the phylogenetic position of Hymenoptera as the sister group of all remaining holometabolous insects (Savard et al. 2006; Peters et al. 2014), the composition of the ancestral genome of Hymenoptera has major implications for understanding the evolution of holometabolous insects and their genomes. Previous studies on Apocrita have shown that the repertoire of immune response genes (Evans et al. 2006; Gadau et al. 2012; Barribeau et al. 2015), of vision genes (opsins) (Henze and Oakley 2015), and the GC content (Standage et al. 2016) of

Hymenoptera genomes are reduced compared with genomes of other insects. A reduction has also been found in the diversity and abundance of transposable elements (TEs), which are key drivers of genome size evolution in insects (Petersen et al. 2019), in social Apocrita (Kapheim et al. 2015). It remains to be investigated, however, whether these traits are characteristic of all Hymenoptera or whether they are specific to Apocrita. Also of interest are the origin and diversification of major royal jelly proteins (MRJPs), which were first discovered in the eponymous royal jelly (Hanes and Šimuth 1992), a honeybee gland secretion fed by young worker bees to developing larvae and triggering queen development (Snodgrass 1925). These proteins are encoded by a varying number of genes (*mrjp* and *mrjp*-like) that are exclusive to Hymenoptera and have been found in all but one of their genomes sequenced thus far (Bonasio et al. 2010; Werren et al. 2010; Nygaard et al. 2011; Smith CR, Smith CD, et al. 2011; Smith, Zimin, et al. 2011; Kupke et al. 2012; Buttstedt et al. 2014; Kapheim et al. 2015; Sadd et al. 2015). The *mrjp-l* genes likely originated from yellow genes (Hanes and Šimuth 1992), which are found across insects, but it is unknown when they originated and started to diversify in Hymenoptera. The current taxonomically biased distribution of genome sequencing data prevents the reliable inference of the ancestral features of Hymenoptera genomes and genomic traits that likely fostered the evolution of parasitoidism.

Here, we present comparative analyses of draft genomes of the ectophytophagous sawfly *Athalia rosae* and the parasitoid sawfly *Orussus abietinus*. *Athalia rosae* (Tenthredinoidea) is a representative of Eusymphyta, which a recent phylogenetic analysis suggests to be the sister lineage of all remaining Hymenoptera (Peters et al. 2017). *Athalia rosae* has retained the ancestral ectophytophagous lifestyle of Hymenoptera and feeds on crucifers (Brassicaceae), of which it is also an important agricultural pest (Sáringer 1974; Abe 1988). The species is readily bred under lab conditions, is currently being established as a model species, and is studied for a wide range of research questions (e.g., in developmental biology, Yamamoto et al. 2004; Sekine et al. 2015; on sex determination, Mine et al. 2017; and on chemical defense, Abdalsamee and Müller 2012). *Orussus abietinus* is a representative of the relatively species-poor group of parasitoid sawflies (also referred to as parasitic wood wasps), consisting exclusively of the family Orussidae. Like other orussids, *O. abietinus* is an ectoparasitoid of xylophagous larvae (beetles and wood wasps) developing in dead wood, a lifestyle considered to likely mirror the ancestral state of parasitoids (Peters et al. 2017). Orussids detect their hosts via vibrational sounding: the female wasps generate vibrations via frequent tapping of the antennae against the wood. The reflecting vibrations (containing information on the presence of host larvae in the wood) are in turn picked up by the forelegs and transmitted through the hemolymph to specialized organs, where they are transduced into nerve impulses (Vilhelmsen et al. 2001). If a host larva is detected, the female orussid lays an egg on or close to the host larva, which the orussid larva feeds upon when hatched (Ahnlund and Ronquist 2002). The anatomy of the orussid larva is simplified compared with those of other sawflies and is more similar to those of Apocrita (Vilhelmsen 2003). For example, orussid larvae lack eyes and legs (as do the larvae of Apocrita and in contrast to the larvae of sawflies) and their antennae and mouthparts are strongly simplified (Vilhelmsen 2003). These morphological characteristics are considered adaptations to a parasitoid lifestyle. Our analyses of the draft genomes of *At. rosae* and *O. abietinus*, including comparisons with those of other Hymenoptera, provide first insights into 1) the composition of the ancestral genome of Hymenoptera, 2) traits related to the transition from phytophagy to parasitoidism, and 3) features that enabled the massive speciation of Apocrita. We also revisit multiple long-standing ideas on hymenopteran genome evolution, the results of which highlight the importance of comprehensive taxonomic sampling in comparative genomics.

Results and Discussion

We sequenced and assembled the genome of the turnip sawfly, *At. rosae*, (Tenthredinidae; a representative of the phytophagous "Symphyta"; fig. 1*A–C*) and the parasitoid sawfly *O. abietinus* (Orussidae; a representative of the parasitoid "Symphyta"; fig. 1*A–C*) at a base coverage depth of 525× and 255×, respectively, from Illumina paired-end and mate-pair libraries using DNA of haploid males (supplementary section II.1, Supplementary Material online). After assembling the reads with ALLPATHS-LG (Gnerre et al. 2011) and scaffolding the resulting contigs using Atlas-Link and Atlas-Fill, the draft genome assemblies of *At. rosae* and *O. abietinus* span 164 and 201 Mb, respectively (fig. 1*D*). The assembly sizes closely match in silico genome size estimates (170 and 247 Mb) inferred from the 17-mer distribution in the Illumina paired-end reads. The two genome assemblies are of high contiguity (522 and 936 scaffolds with N50 of 1.37 and 2.37 Mb, respectively) compared with other Hymenoptera draft genome assemblies (supplementary file S1, Supplementary Material online). Assessments of gene space coverage using the Arthropoda gene set of Benchmarking Universal Single-Copy Orthologs (BUSCO; Simão et al. 2015) further revealed that the genome assemblies encompass the majority (96% and 93%) of the expected protein-coding genes (supplementary section II.2.1, Supplementary Material online). The assemblies are close in size to that of the wheat stem sawfly, *Ce. cinctus* (Cephidae; 205 Mb; Hanrahan and Johnston 2011; Robertson et al. 2018), and fall within the lower range of the known genome sizes of Hymenoptera (98 Mb–1.3 Gb; Ardila-Garcia et al. 2010; Tavares et al. 2010; Gregory 2018).

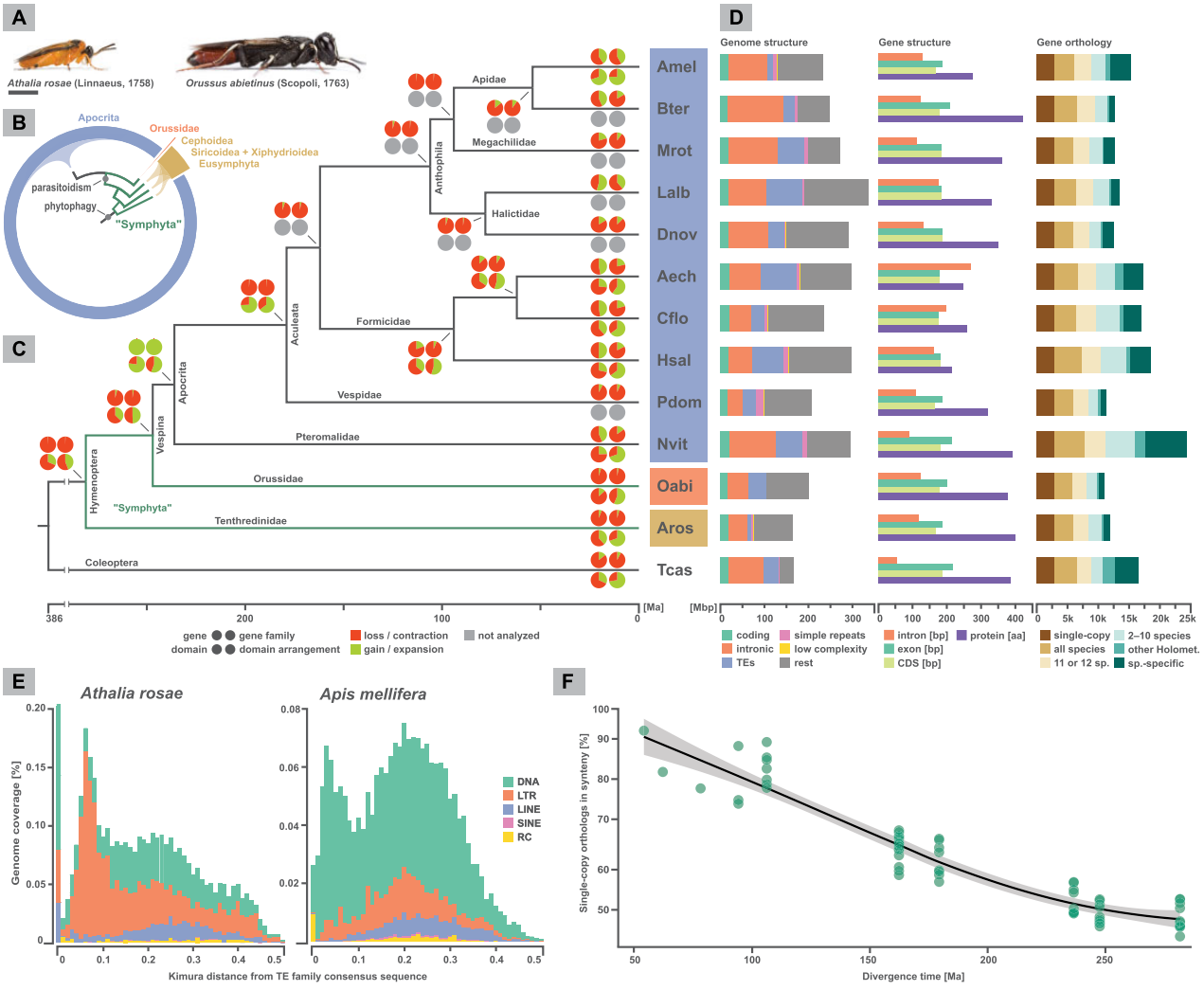

Fig. 1.—Hymenoptera genome evolution. (*A*) Adult males of *Athalia rosae*[a] and *Orussus abietinus*. Scale bar: 2.5 mm. (*B*) Number of described species (Apocrita: 144,593; Orussidae: 82; "Symphyta" excl. Orussidae: 7,983) of, relationships of, and ecological transitions in Hymenoptera (Aguiar et al. 2013; Peters et al. 2017). (*C*) Ratio of gain and loss of genes, domains, and domain arrangements, as well as ratio of gene families that experienced expansions or contractions. Gene and gene family evolution were analyzed by applying the maximum-likelihood optimality criterion, a single coupled birth and death rate, and using the divergence time estimates and phylogenetic relationships inferred by Peters et al. (2017). Domain and domain arrangement evolution were analyzed by applying the maximum parsimony optimality criterion. (*D*) Absolute number of nucleotides occupied by genomic components (left column), median length of various gene structure parameters (center column), and gene orthology in the genome of each species (right column; unit = number of genes). (*E*) Divergence distribution of transposable element (TE) copies in the genome of *At. rosae* and that of *Apis mellifera*, estimated from the Kimura distance of the nucleotide sequence of each TE copy to the TE family nucleotide consensus sequence. (*F*) Loss of synteny over time in the genomes of 12 Hymenoptera, inferred from the proportion of 3,983 shared single-copy orthologs (SCOs) retaining the same neighboring SCO, relative to the divergence time, in all possible pairwise comparisons. The curve represents the smoothed conditional mean. aa, amino acids; bp, base pairs; CDS, coding sequence; LINE, long interspersed nuclear element; LTR, long terminal repeats; Ma, million years ago; RC, rolling circle transposons; SINE, short interspersed nuclear element; TE, transposable elements; Aech, *Acromyrmex echinatior*; Amel, *Apis mellifera*; Aros, *A. rosae*; Bter, *Bombus terrestris*; Cflo, *Camponotus floridanus*; Dnov, *Dufourea novaeangliae*; Hsal, *Harpegnathos saltator*; Lalb, *Lasioglossum albipes*; Mrot, *Megachile rotundata*; Nvit, *Nasonia vitripennis*; Oabi, *Orussus abietinus*; Pdom, *Polistes dominula*; Tcas, *Tribolium castaneum*. All photographs by Oliver Niehuis, with assistance from Thomas Pauli and Ralph S. Peters. [a]Note that while the photograph shows a male of the nominate form, we sequenced and report the genome of the Eastern Palearctic subspecies *At. rosae ruficornis*.

In fact, the genome of *At. rosae* is the smallest of all Hymenoptera sequenced so far. The two sawfly genomes have a higher GC content than most apocritan genomes (sawflies: 41% and 45%; Apocrita: median 37%; supplementary section II.4.2, Supplementary Material online). This is consistent with the hypothesis that the low GC content of Apocrita genomes represents a derived state, possibly caused by high recombination rates associated with GC-

biased gene conversion (Wilfert et al. 2007; Niehuis et al. 2010; Kent et al. 2012). However, the cause and effect relationship of recombination rate and GC content remains to be disentangled.

Copy Number and Amino Acid Sequence of Conserved Genes and Gene Families Substantiate the High Quality of the Sawfly Draft Genomes

The evolution of the hymenopteran gene repertoire was studied in detail by manually annotating >1,000 protein-coding genes and noncoding RNAs in each of the two sawfly genomes. We found a wide range of genes and gene families to be conserved in amino acid sequence and copy number across Hymenoptera, consistent with a priori expectations, and confirming the high coverage of the sawfly genomes by the draft assemblies. Manually annotated and studied genes and gene families include ncRNAs, potentially laterally transferred genes, MRJPs, storage proteins, developmental genes, insulator proteins, DNA methyltransferases, silk proteins, elongases, desaturases, opsins, metallopeptidases, heat shock proteins, aquaporins, cuticular proteins, cysteine peptidases, candidate venom proteins, neuropeptides, protein hormones, biogenic amines, and their G-protein-coupled receptors, as well as genes related to chemoreception, immune response, autophagy, dosage compensation, RNA interference, antioxidants, sex determination, and oxidative phosphorylation. A full description and discussion of each of these genes or gene families is given in the Supplementary Material online (supplementary sections II.4.4 and II.5.1–25, Supplementary Material online).

TE Content and Activity

Diversity and abundance of TEs, and consequently genome size, have been found to negatively correlate with the degree of social complexity in Apocrita (Kapheim et al. 2015). This is possibly a consequence of high recombination rates and decreased exposure to parasites and pathogens in eusocial species (Kapheim et al. 2015). We found the relative TE content in genomes of Hymenoptera, identified with RepeatModeler (Smit and Hubley 2015) and RepeatMasker (Smit et al. 2015), to strongly correlate with genome size (Pearson's product-moment correlation $r = 0.8$, $P = 0.003$; supplementary section II.3.5, Supplementary Material online) and to range from 4.7% (11.0 Mb) in the honeybee (*Apis mellifera*) to 27.4% (81.5 Mb) in the leaf-cutting ant (*Acromyrmex echinatior*) (fig. 1*D* and supplementary file S4, Supplementary Material online). TE sequence divergence analysis, based on intrafamily Kimura 2-parameter distances, indicates recent peaks in TE activity, largely caused by DNA elements, in most Hymenoptera genomes (supplementary fig. S7, Supplementary Material online). Interestingly, the *At. rosae* genome shows a TE content (5.1%) and TE activity spectrum that is, with the exception of a very recent burst of TEs, similar to that of the honeybee (fig. 1*E*). These results are intriguing, since they demonstrate that low TE content and overall low TE activity in Hymenoptera are not restricted to genomes of eusocial species and that consequently other ultimate factors seem to govern TE content evolution.

Apocrita Possess More Genes with Reduced Gene Structure Complexity than Sawflies

The automated MAKER protein-coding gene annotation pipeline (Cantarel et al. 2007) predicted 11,894 and 10,959 genes in the draft genomes of *At. rosae* and *O. abietinus*, respectively. The numbers of genes predicted in the two sawfly draft genomes are lower than the official gene counts of most other published Hymenoptera draft genomes (fig. 1*D*; Branstetter et al. 2018), but closely match the reported numbers of protein-coding genes in the draft genomes of *Ce. cinctus* (11,206; Robertson et al. 2018) and the European paper wasp, *Polistes dominula* (fig. 1*D*; Standage et al. 2016). However, comparing features of the predicted protein-coding genes across species using COGNATE (Wilbrandt et al. 2017) revealed that the total amount of protein-coding DNA in the two sawfly genomes (19.9 Mb in *At. rosae* and 17.7 Mb in *O. abietinus*) fits well into the known range of the metric in Hymenoptera (16–20 Mb; fig. 1*D*) and that the total amount of protein-coding DNA varies less than the number of genes across the published draft genomes of Hymenoptera. Proteins of the two sawflies are among the longest in Hymenoptera (median: 406 amino acids in *At. rosae* and 384 amino acids in *O. abietinus*; fig. 1*D*). The protein length increase results from a larger median number of exons (5.0; note that the sizes of exons in the sawfly draft genomes do not differ markedly from the average across Hymenoptera; supplementary section II.4.2, Supplementary Material online), compared with Apocrita (4.0).

Gene Order Is Constrained in Hymenoptera

Gene order is subject to change over the course of evolution due to recombination and rearrangement. Because genome-wide recombination rates vary substantially between Hymenoptera, with eusocial species likely exhibiting the highest rates (Wilfert et al. 2007), the rate of microsynteny (gene order conservation) decay is also expected to differ between lineages. Yet, previous studies have found extensive conservation of gene order across insects (Engström et al. 2007). Using protein divergence as a proxy for time, a linear decay of microsynteny over time has been found in insect genomes (Zdobnov and Bork 2007). Capitalizing on recently published Hymenoptera divergence time estimates (Peters et al. 2017) and exploring a more extensive taxon sampling within Hymenoptera, including the two sawflies presented here, we investigated microsynteny decay of conserved single-copy orthologs (SCOs) in this insect order. Comparing the fraction of SCOs that retain the same neighboring SCO in

pairwise comparisons between species in relation to the divergence times of each species pair using a custom Perl script (included as supplementary file S39, Supplementary Material online) revealed a close to linear loss of synteny over time (fig. 1*F*). The highest degree of synteny conservation was detected between the most recently diverged lineages (e.g., > 90% between honeybee and leafcutter bee; supplementary file S38, Supplementary Material online), irrespective of whether these lineages are eusocial or not. In fact, we did not observe an increase of genome shuffling in eusocial Apocrita. However, contrary to what was previously reported by Zdobnov and Bork (2007), we found a decrease in the rate of synteny loss across divergence times that span >240 Myr (fig. 1*F*). This retention of microsynteny over large evolutionary distances points to the presence of functional constraints on the preservation of local genomic structures or low rates of nonhomologous recombination and rearrangement. Functional annotation of genes remaining in microsynteny, using Gene Ontology terms, revealed significant enrichment ($P < 0.05$; weighted Fisher's test and hypergeometric test) of a number of terms related to cell cycle and signaling, cellular and organelle organization, as well as development (supplementary file S2, Supplementary Material online). Notably, we found consistent enrichment of Wnt and Notch signaling, both of which are vital and complex pathways in embryonic development and tissue differentiation. A specific example of a conserved gene order was also revealed by manual annotation of opsin genes (supplementary section II.5.24, Supplementary Material online): we uncovered a close linkage of the long-wave sensitive (LWS) 1 and LWS 2 opsins, which was previously considered unique to the honeybee (Bao and Friedrich 2009), in the genomes of the two sawflies and of ten additional hymenopterans (interlocus distance: -6–7,583 bp; supplementary file S35, Supplementary Material online). The conserved LWS1/2 linkage thus represents an ancestral feature of all Hymenoptera and suggests the presence of a *cis*-regulatory constraint, preventing the loss of synteny between these genes.

Hymenoptera Gene and Protein Domain Repertoires Display a Reductive Mode of Evolution

A previous study reported that more genes were gained than lost in the evolution of protein-coding gene families in Hymenoptera (Rappoport and Linial 2015). Here, we analyzed the evolution of gene families inferred from OrthoDB (Zdobnov et al. 2017) using the CAFE software (Han et al. 2013) and exploiting recently published divergence time estimates of Hymenoptera (Peters et al. 2017). We additionally identified protein domains as well as protein domain arrangements and inferred their respective losses and gains across the Hymenoptera tree applying the Fitch parsimony optimality criterion. In contrast to the study of Rappoport and Linial (2015), we found a pronounced pattern of reduction of genes, gene families, and protein domains during the evolution of this insect order, with more losses than gains at most nodes (fig. 1*C* and supplementary file S41 and section II.4.3, Supplementary Material online). The pattern is contrasted by a large number of new protein domain arrangements uncovered at each node (supplementary fig. S11, Supplementary Material online), with more new arrangements than lost arrangements (fig. 1*C*). This result is consistent with the idea that domains can be reused and shuffled at a higher rate than new domains can emerge (Lees et al. 2016; Moore and Bornberg-Bauer 2012). Ultimately, reuse of functional units might compensate for the predominant trend of gene and domain loss as well as for gene family contractions (Lees et al. 2016).

MRJPs Were Already Synthesized by the Last Common Ancestor of Hymenoptera

MRJPs are an important component of the honeybee's royal jelly, a gland secretion fed to developing larvae that determines the differential development of queens and workers (Snodgrass 1925). MRJPs and MRJP-like encoding genes have only been known to occur in Apocrita, presumably having evolved from a tandem-duplication of the Yellow-family gene *y-e3* and subsequently expanded in multiple lineages (Drapeau et al. 2006; Buttstedt et al. 2014). Revising this scenario, manual annotation of MRJPl-encoding genes uncovered a single gene in the genome of *O. abietinus* and ten genes in the genome of *At. rosae* (fig. 2*A* and supplementary section II.5.5 and fig. S18, Supplementary Material online). The presence of a single *mrjpl* in the genome of *O. abietinus* is consistent with the hypothesis of a single ancestral *mrjpl* in Apocrita (Drapeau et al. 2006), but with its origin already in a stem lineage of all Hymenoptera. The evolutionary origin of *mrjpls* (> 281 Ma) is thus much older than previously thought. Phylogenetic analysis recovered *mrjpls* as sister group of the Yellow-gene *y-x2* and not of the Yellow-gene *y-e3* (fig. 2*A* and supplementary fig. S19, Supplementary Material online), despite a higher similarity of *mrjpls* in intron–exon structure with the latter (fig. 2*B* and supplementary section II.5.5, Supplementary Material online). This phylogenetic relationship also received statistically significant support ($P = 0.0048$; approximately unbiased topology test). The close relationship of *mrjpls* with *y-x2* is especially surprising, given that *y-x2* is spatially distantly located from the *yellow* gene cluster containing the *mrjpls* (e.g., in the genomes of *Ap. mellifera* and *Nasonia vitripennis*, they occur on different chromosomes; Drapeau et al. 2006; Buttstedt et al. 2014; in the genomes of *At. rosae* and *O. abietinus*, they occur on different scaffolds).

Hymenoptera Are Characterized by a Small Repertoire of Conserved Immune Genes

The canonical immune response-related gene repertoire involved in recognition and signaling pathways (immune genes)

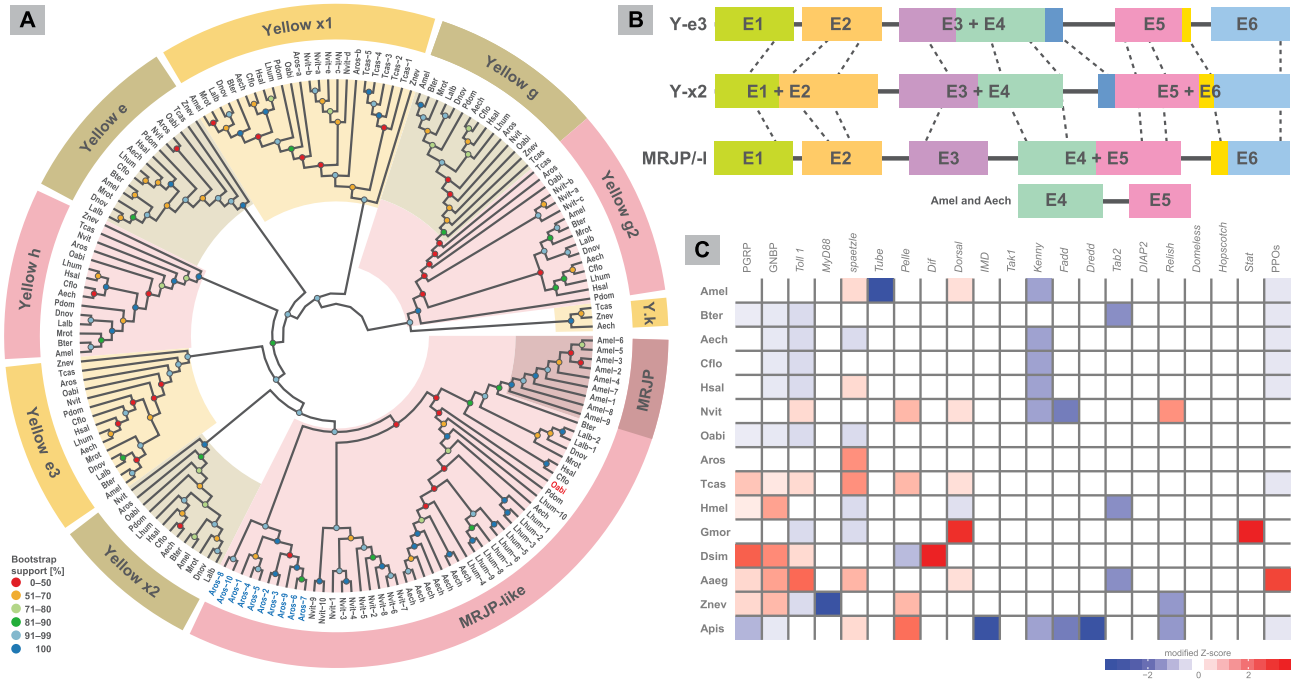

Fig. 2.—Evolution of hymenoptera yellow, MRJP/-like, and immune response-related genes. (*A*) Relationships of hymenoptera yellow, major royal jelly protein (MRJP), and MRJP-like (MRJPl) amino acid sequences, inferred under the maximum-likelihood optimality criterion, modeling invariable sites, and approximating site-rate variation with a discrete gamma distribution. Branch support is estimated from 1,000 nonparametric bootstrap replicates. MRJP and MRJPl proteins of *Athalia rosae* and *Orussus abietinus* are highlighted in blue and red, respectively. (*B*) Gene structure comparison of *mrjp* and *mrjpl* genes and of two candidate sister group *yellow* genes, *y-e3* and *y-x2*. Dashed lines indicate shared amino acid motifs conserved among species within each gene and between genes (supplementary section II.5.5, Supplementary Material online). Gene and motif lengths not to scale. (*C*) Heat map visualizing copy number variation in immune response-related genes between species. Modified *Z*-scores indicate the deviation from the median of each gene by SD units. Aaeg, *Aedes aegypti*; Aech, *Acromyrmex echinatior*; Amel, *Apis mellifera*; Apis, *Acyrthosiphon pisum*; Aros, *Athalia rosae*; Bter, *Bombus terrestris*; Cflo, *Camponotus floridanus*; Dnov, *Dufourea novaeangliae*; Dsim, *Drosophila simulans*; Gmor, *Glossina morsitans*; Hmel, *Heliconius melpomene*; Hsal, *Harpegnathos saltator*; Lalb, *Lasioglossum albipes*; Lhum, *Linepithema humile*; Mrot, *Megachile rotundata*; Oabi, *Orussus abietinus*; Pdom, *Polistes dominula*; *Nvit*, *Nasonia vitripennis*; Tcas, *Tribolium castaneum*; Znev, *Zootermopsis nevadensis*.

of eusocial Hymenoptera was initially described as extremely reduced compared with the mostly conserved repertoire of solitary insects (Evans et al. 2006; Gadau et al. 2012). However, a more recent study suggested that a reduced immune gene repertoire might be a shared trait of Apocrita and is not strictly correlated with a eusocial lifestyle (Barribeau et al. 2015). Using profile hidden Markov models built from reference amino acid sequences of immune genes to scan the predicted proteins of Hymenoptera and a selected set of other insects, we found the numbers of immune genes to be largely similar among all investigated species of Hymenoptera (28–36 genes; fig. 2C and supplementary table S25, Supplementary Material online), although some lineages are characterized by the lack of specific genes (e.g., the IMD pathway gene *Kenny* is absent in several Aculeata). Although the genome of *At. rosae* has the largest number of identified response-related genes among Hymenoptera, our data do not show a clear trend between immune gene repertoire reduction and eusocial lifestyle. On the contrary, we found 32 immune genes in the genome of the eusocial honeybee, but only 29 in that of the solitary *O. abietinus* (supplementary table S25, Supplementary Material online). We also found that Hymenoptera are characterized by an overall small number of immune genes (median: 30) relative to other insects (median: 38; supplementary table S25, Supplementary Material online). The reduced number of immune genes in Hymenoptera is thus likely not related to the evolution of eusociality, nor is it a characteristic of Aculeata, but rather represents the ancestral condition in Hymenoptera. However, the reduced repertoire of recognition and signaling pathway genes, which are mostly conserved across solitary insects, in Hymenoptera does not necessarily imply a reduced immune response. A study investigating de novo infection response genes in *N. vitripennis* identified a large repertoire of new genes involved in the immune response, many of which were taxonomically restricted and rapidly evolving (Sackton et al. 2013). It remains to be tested if and how these novel immune response-related genes compensate for the reduction of the immune gene repertoire and also whether such a compensation has evolved in other Hymenoptera.

Loss of a Vision Gene Coincides with Transition to a Parasitoid Lifestyle

Light sensing is primarily mediated by the opsin gene family of G protein-coupled transmembrane receptors. Apocrita are known to have four rhabdomeric-type opsins (r-opsins) of three wavelength-specific subfamilies: one member of the short-wavelength UV-sensitive (SWS-UV) r-opsin subfamily, one member of the blue-sensitive (SWS-B) r-opsin subfamily, and two members of the long wavelength-sensitive (LWS) r-opsin subfamily, introduced above as LSW1 and LSW2 opsins (Velarde et al. 2005; Wakakuwa et al. 2005; Henze and Oakley 2015). These r-opsins are differentially expressed in the photoreceptors of the compound eye retina and the ocelli (Velarde et al. 2005). The honeybee has also been shown to possess a fifth opsin, a member of the ciliary opsin gene family (c-opsin), which is expressed in two small cell clusters of the brain, likely mediating extraretinal light sensing (Velarde et al. 2005). Using known opsin amino acid sequences as references, we identified and manually annotated all four retinal opsins that had previously been found in Hymenoptera in the genomes of the two sawflies (fig. 3*A* and supplementary table S27, Supplementary Material online). This revealed that the molecular underpinnings underlying trichromatic compound eye vision, which has been documented by comparative physiological studies in the Hymenoptera (Peitsch et al. 1992), is highly conserved in the order. Furthermore, we found that the c-opsin is also present in the *At. rosae* genome (fig. 3*A*) and that the *At. rosae* genome is unique among Hymenoptera in containing a sixth opsin, Rh7 (fig. 3*A*). The Rh7 opsin is deeply conserved in arthropods (Senthilan and Helfrich-Förster 2016), but is not found in other Hymenoptera, suggesting that this opsin subfamily was lost in the stem lineage of Orussidea and Apocrita. In *Drosophila*, Rh7 opsin has been found to be expressed in the brain and is involved in the entrainment of the circadian activity rhythm by light (Ni et al. 2017). However, Rh7 opsin is also expressed in the photoreceptor cells of a mosquito species (Hu et al. 2014). Thus, besides identifying *At. rosae* as the opsin homolog-richest hymenopteran species at this point, these findings revealed that the transition from phytophagy to a parasitoid lifestyle in Hymenoptera was accompanied by a reduction of the opsin gene repertoire. This could be related to the extreme regression of the larval visual system as ancestral parasitoid larvae are thought to have developed in wood and were thus not exposed to sunlight (Vilhelmsen and Turrisi 2011).

Dietary Transition and Specialization Have Not Resulted in Change of Metabolic Capabilities

Phytophagous sawfly larvae, being mobile in the environment, can utilize multiple host plants or prey. In contrast, parasitoid larvae are restricted to a single host and the finite resources contained within this host (Jervis et al. 2008). To alleviate the severely limited resources available to each parasitoid larva (Slansky 1986; Jervis et al. 2008), some highly specialized parasitoids manipulate their host to increase nutrient availability (Vinson and Iwantsch 1980). As a consequence, many of these parasitoids have in turn lost the ability to synthesize these nutrients (e.g., lipids), possibly through the loss of synthesis pathway genes (Visser et al. 2010). However, the genomic changes of the metabolic gene repertoire associated with the transition from phytophagy to parasitoidism and from generalist to specialist parasitoid have not been comprehensively characterized in Hymenoptera. We functionally annotated the predicted proteins of the phytophagous sawfly (*At. rosae*), the generalist parasitoid sawfly (*O. abietinus*; host spectrum reviewed by Ahnlund and Ronquist 2002), and the highly specialized parasitoid wasp (*N. vitripennis*; host spectrum discussed by Peters 2010 and Desjardins et al. 2010) using the CycADS pipeline (Vellozo et al. 2011). We then inferred and compared metabolic pathways in the three species through a combination of a Pathway Tools (Karp et al. 2016) analysis and manual curation of the results. We found fewer genes with predicted metabolic functions in *At. rosae* (4,090; supplementary section II.5.10, Supplementary Material online) and *O. abietinus (*3,827) than in *N. vitripennis* (4,928). Despite these differences, we found a high level of congruence in the enzyme repertoire and in the metabolic pathways between all three species (fig. 3*B* and *C*). Surprisingly, the comparison of the predicted functions of the inferred enzymes and pathways did not reveal differences that can readily be attributed to dietary transitions or host specialization. The lack of any detectable reduction in the metabolic gene repertoire of the two parasitoids can possibly be explained by the propensity of adult parasitoid Hymenoptera to consume pollen, nectar, and plant tissue (Jervis et al. 1993), for which the ancestral metabolic gene repertoire is still required. The dietary transitions and specializations during the evolutionary history of Hymenoptera might consequently not have resulted in the complete loss of metabolism-related gene families, but might have instead caused a reduction in the copy number of genes, as was shown in mammals (Kim et al. 2016), or changes in gene expression and enzyme efficiency. Consistent with this idea, the manual annotation of genes that are likely related to the ability of *At. rosae* to deal with the chemical defenses of its host plant also reflects this pattern, revealed a reduced copy number of two candidate gene families in carnivorous and secondarily phytophagous species relative to ancestrally phytophagous species (supplementary section II.5.9, Supplementary Material online). A partial repertoire reduction could thus explain how the ability to synthesize lipids has re-evolved multiple times in parasitoid wasps (Visser et al. 2010).

Storage Protein Evolution Possibly Facilitated Transition to Parasitoidism

The efficient utilization of limited host resources by larvae and the allocation of these resources to the adult stage are

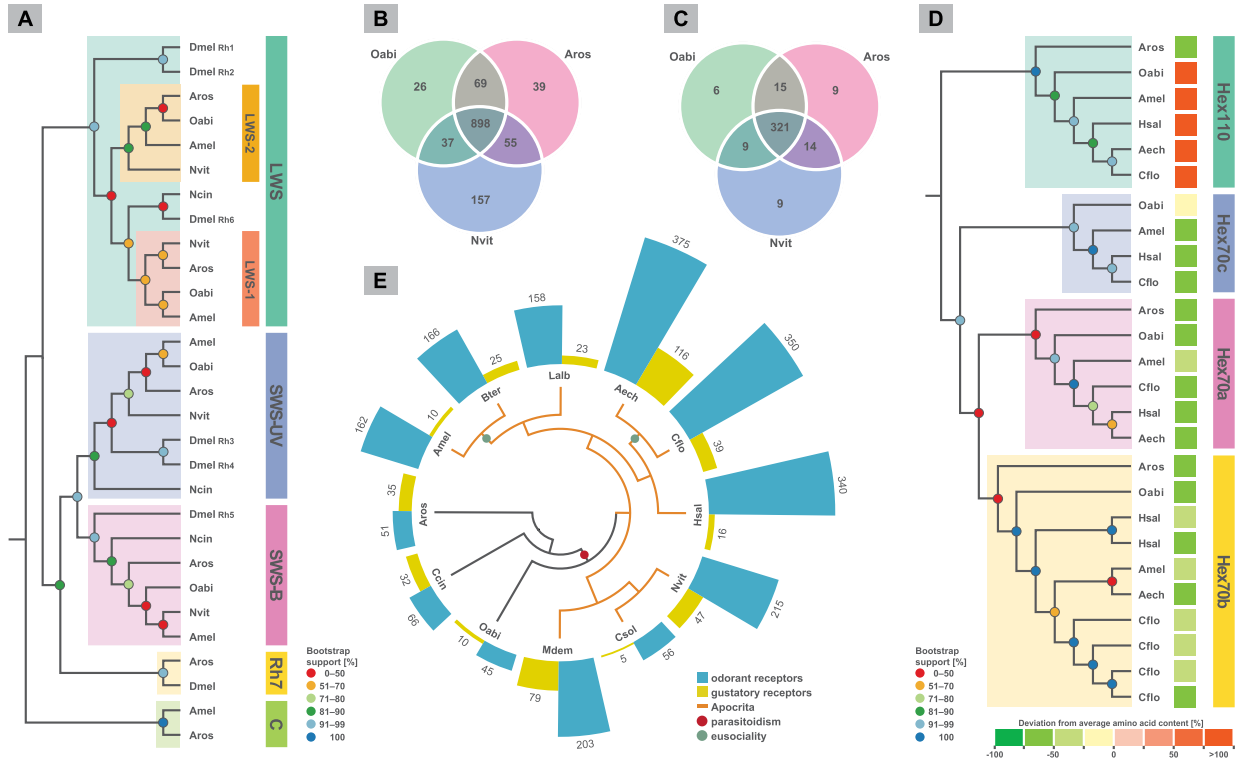

Fig. 3.—Hymenoptera vision gene, metabolic, hexamerin, and chemoreceptor repertoires. (*A*) Phylogenetic relationships of Hymenoptera, *Nephotettix cincticeps* (Hemiptera), and *Drosophila* opsin genes inferred under the maximum-likelihood optimality criterion. Branch support is estimated from 500 nonparametric bootstrap replicates. (*B*) Number of unique and shared enzymes (Enzyme Commission numbers) in the proteomes of *Athalia rosae*, *Orussus abietinus*, and *Nasonia vitripennis*. (*C*) Number of unique and shared metabolic pathways identified in the proteomes of *At. rosae*, *O. abietinus*, and *N. vitripennis*, inferred from enzyme and gene ontology annotations. (*D*) Phylogenetic relationships of Hymenoptera hexamerins inferred under the maximum-likelihood optimality criterion. Branch support is estimated from 1,000 nonparametric bootstrap replicates. Colors indicate deviation of the amino acid glutamine (Q) from the average amino acid content in percent (%). (*E*) Copy number variation of odorant and gustatory receptor gene repertoires among Hymenoptera. Data referring to *At. rosae* and *O. abietinus* are taken from the present study, those of all remaining species from literature (Robertson and Wanner 2006; Robertson et al. 2010; Zhou et al. 2012, 2015; Sadd et al. 2015; Robertson et al. 2018). Only full-length proteins comprising at least 350 amino acids were considered. Phylogenetic relationships taken from the study by Peters et al. (2017). Aech, *Acromyrmex echinatior*; Amel, *Apis mellifera*; Aros, *Athalia rosae*; Bter, *Bombus terrestris*; Ccin, *Cephus cinctus*; Cflo, *Camponotus floridanus*; Csol, *Ceratosolen solmsi*; Dmel, *Drosophila melanogaster*; Hsal, *Harpegnathos saltator*; Lalb, *Lasioglossum albipes*; Mdem, *Microplitis demolitor*; Ncin, *Nephotettix cincticeps*; Nvit, *Nasonia vitripennis;* Oabi, *Orussus abietinus*.

essential to the reproductive success of parasitoids (Jervis et al. 2008). Most Apocrita possess four hexamerin storage proteins (Hex70a–c and Hex110; Cristino et al. 2010; Martins et al. 2010), which provide energy and amino acids during non-feeding periods (Hagner-Holler et al. 2007). Utilizing reference amino acid sequences of known hexamerins, we identified and manually annotated all four previously known hexamerins of Hymenoptera in the genome of *O. abietinus* and all but Hex70c in the genome of *At. rosae* (fig. 3*D*). Comparing the amino acid content of Hymenoptera hexamerins, we found a unique and substantial increase of glutamine content (>100% increase relative to the average)—which is important in the management of nitrogen in insects (Weihrauch et al. 2012)—in the Hex110 protein of *O. abietinus* and of all Apocrita (fig. 3*D*). This change might have evolved in response to the increased nitrogen content in animal tissues relative to plant matter (Mattson 1980). Thus, the emergence of an additional hexamerin storage protein (Hex70c) and the increased level of glutamine in Hex110 in the stem lineage of *O. abietinus* and Apocrita possibly facilitated the transition from a herbivorous to a parasitoid lifestyle.

Odorant and Gustatory Receptors Were Likely Key Factors for the Diversification of Apocrita

Chemosensation receptors are paramount for vital insect behaviors, such as host detection in parasitoid wasps (Steidle and Schöller 1997). Hymenoptera detect most chemical compounds with transmembrane proteins of the odorant receptor (OR) and of the gustatory receptor (GR) multigene families. These families are very diverse in Apocrita and especially so in lineages with eusocial species (fig. 3*E*), where they possibly facilitated the evolution of eusociality by enabling kin selection (Zhou et al. 2012). We identified and manually

annotated odorant and gustatory receptors in the two sawfly genomes utilizing the antennal transcriptomes of each species and a set of reference amino acid sequences of the corresponding proteins in other Hymenoptera. In agreement with a recent study on *Ce. cinctus* (Robertson et al. 2018), we found considerably fewer GR- and OR-coding genes in the genomes of the two sawflies *At. rosae* and *O. abietinus* than in those of Apocrita (fig. 3*E* and supplementary section II.5.3, Supplementary Material online). In addition, our data indicate that multiple OR and GR gene lineages present in the genomes of the herbivorous sawflies *At. rosae* and *Ce. cinctus* were lost during the evolution of parasitoidism in the last common ancestor of *O. abietinus* and Apocrita (fig. 3*E* and supplementary section II.5.3, Supplementary Material online). The large OR and GR gene repertoires of Apocrita are the result of subsequent and multiple independent expansions of those OR and GR gene lineages that were retained during the evolution of parasitoidism (supplementary section II.5.3, Supplementary Material online). Most intriguingly, the 9-exon OR subfamily, which has been implicated in the detection of cuticular hydrocarbons and is particularly expanded in eusocial species (up to 139 in the red harvester ant, *Pogonomyrmex barbatus*; Smith CR, Smith CD, et al. 2011; Zhou et al. 2012, 2015; Pask et al. 2017) is represented by only one copy in each of the sawfly genomes (supplementary fig. S16, Supplementary Material online). The expansion of the OR and GR gene repertoires in Apocrita likely improved the chemoreception abilities of apocritans and could thus have been a key factor in the evolutionary success of this group. Specifically, the improved chemoreception abilities may have facilitated the formation of new ecological niches by enabling efficient detection and differentiation of novel hosts in diverse habitats. Encountering new hosts is key for specialization (Schmid-Hempel 2011), which in turn enables parasitoids to evolve faster and adapt more readily to the host defense mechanisms (Kawecki 1998). Consistent with this idea, the species-poor parasitoid orussids identify potential hosts in wood via vibrational sounding (Vilhelmsen et al. 2001), which likely provides far fewer possibilities for host specialization than chemoreception. Finally, we found two of the GR genes in the genomes of *At. rosae* and *Ce. cinctus* to be orthologous to CO_2 receptor genes of *Drosophila* (Jones et al. 2007; Kwon et al. 2007; Robertson et al. 2018). The presence of candidate CO_2 receptor genes in the genomes of phytophagous sawflies, in contrast to their absence in the genomes of the parasitoid sawfly and Apocrita, could thus indicate the functional involvement of the encoded receptors in host plant detection.

Conclusions

The results from our comparative analyses of the *At. rosae* and *O. abietinus* genomes call several previously widely held assumptions regarding characteristics and the evolution Hymenoptera genomes in to question. It has been stated, for example, that Hymenoptera genomes are characterized by a low GC content (Standage et al. 2016; Branstetter et al. 2018). Considering the phylogenetic relationships of the investigated species, the high GC content of sawfly genomes does not represent a simple exception from a rule, but suggests that a low GC content is a derived state of only a subordinate group of Hymenoptera, the Apocrita. Contrariwise, we uncover genomic attributes previously considered derived characteristics of highly specialized lineages (e.g., bees) to actually represent Hymenoptera ground plan features (e.g., presence of MRJPls and a reduced immune response gene repertoire). We also provide novel insights into genomic factors that may have facilitated the evolutionary success and the tremendous diversification of parasitoid and eusocial Apocrita (e.g., changes in storage protein and chemosensory receptor repertoires). The results of our study highlight the importance of taxonomic sampling for inferring ground plan characteristics of an organismal group. They furthermore lay the foundation for a variety of future lines of research (e.g., on the ancestral function of MRJPs and the possible fitness benefits of the CO_2 receptors) by providing a valuable resource for comparative studies in the mega-diverse insect order Hymenoptera, which encompasses economically (Quicke 1997; Grimaldi and Engel 2005) and medically relevant (Moreno and Giralt 2015) species as well as important model organisms (Weinstock et al. 2006; Werren et al. 2010; Branstetter et al. 2018).

Materials and Methods

Samples and Extractions

All samples of *At. rosae ruficornis* Jakovlev, 1888 were derived from a strain maintained for >15 years, with occasional introductions of individuals from natural populations, in the laboratory of M. Hatakeyama (National Institute of Agrobiological Sciences NARO, Tsukuba, Japan). Total genomic DNA was extracted from adult haploid males originating from a single virgin female using the Gentra Puregene Tissue Kit (Qiagen, Hilden, Germany) and following the manufacturers' protocol. Total RNA was extracted from the whole body of 1) two adult males and 2) two adult females using the RNeasy Mini Kit (Qiagen, Hilden, Germany) as well as from the antennae of 3) 45 adult females and 4) 56 adult males using the RNeasy Micro Kit (Qiagen, Hilden, Germany) and following the manufacturers' protocol. Antennae from a given sex were pooled for RNA extraction. Samples of *O. abietinus* (Scopoli, 1763) were derived from a natural population of the species in the vicinity of Darmstadt (Hesse, Germany). Total genomic DNA was extracted from the mesosoma and the metasoma of two adult males using the DNeasy Blood and Tissue Kit (Qiagen GmbH, Hilden, Germany) and following the manufacturers' protocol. Total RNA was extracted from 1) the mesosoma and

the metasoma of an adult male using Tri-Reagent (Sigma-Aldrich, Steinheim, Germany), from 2) a whole adult female using the NucleoSpin RNA II Kit (Macherey and Nagel, Düren, Germany), and from 3), the antennae of ten adult males using RNeasy Micro Kit (Qiagen, Hilden, Germany) following the manufacturers' protocols.

Genome and Transcriptome Sequencing

We applied a whole-genome shotgun sequencing approach and prepared and sequenced four libraries of nominal insert sizes of 180 bp, 500 bp, 2 kb (only *At. rosae*), 3 kb, and 8–10 kb. For sequencing the *At. rosae* genome, the 180- and 500-bp paired-end libraries and the 2-kb mate-pair library were prepared from DNA isolated from a single male each, whereas the 3- and 8- to 10-kb mate-pair libraries were prepared using DNA from four and 14 pooled males, respectively. For sequencing the *O. abietinus* genome, the 180-bp, 500-bp, and 3-kb libraries were prepared from DNA extracted from a single adult male wasp, whereas the 8- to 10-kb mate-pair library was prepared using pooled DNA from two adult male wasps. To prepare the 180- and 500-bp libraries, we used a gel-cut paired-end library protocol. Briefly, 1 μg of the DNA was sheared using a Covaris S-2 system (Covaris, Inc. Woburn, MA) using the 180- and 500-bp program, respectively. Sheared DNA fragments were purified with Agencourt AMPure XP beads, end-repaired, dA-tailed, and ligated to Illumina universal adapters. After adapter ligation, DNA fragments were further size selected by agarose gel separation and were subsequently PCR-amplified with 6–8 amplification cycles using the Illumina P1 and Index primer pair and the Phusion High-Fidelity PCR Master Mix (New England BioLabs, Ipswich, MA). The final library was purified using Agencourt AMPure XP beads, and the library's quality was assessed with an Agilent Bioanalyzer 2100 (DNA 7500 Kit) by determining the fragment size distribution. Long mate-pair libraries with 2-, 3- and 8- to 10-kb insert sizes were constructed according to the manufacturer's protocol (Mate Pair Library v2 Sample Preparation Guide Art No. 15001464 Rev. A PILOT RELEASE). Briefly, 5 μg (when preparing the 2- and the 3-kb insert size libraries) or 10 μg (8- to 10-kb insert size library) of genomic DNA was sheared to the desired fragment size with the aid of a Hydroshear (Digilab, Marlborough, MA). The obtained fragments were subsequently end-repaired and biotinylated. Fragment sizes between 1.8 and 2.5 kb (2-kb library), between 3.0 and 3.7 kb (3-kb library), and between 8 and 10 kb (8- to 10-kb library) were extracted from a 1% low-melting agarose gel and then circularized by blunt-end ligation. The size-selected circular DNA fragments were then sheared to fragment sizes of 400 bp (Covaris S-2), the fragments were subsequently purified using Dynabeads M-280 Streptavidin Magnetic Beads, end-repaired, dA-tailed, and ligated to Illumina PE sequencing adapters. DNA fragments with adapter molecules on both ends were amplified for 12–15 cycles with Illumina P1 and Index primers. Amplified DNA fragments were purified with Agencourt AMPure XP beads. Quantification and size distribution of the final library were determined before sequencing as described above. All sequencing was performed on Illumina HiSeq2000 sequencers, which generated 100-bp paired-end reads. Using a genome size estimate of 170 Mb as baseline (see supplementary section II.1.3.1, Supplementary Material online), we sequenced the five *At. rosae* libraries (i.e., 180 bp, 500 bp, 2 kb, 3 kb, and 8–10 kb) to base coverage depths of 240×, 62×, 57×, 109×, and 57×, respectively. Using a genome size estimate of 247 Mb as baseline (supplementary section II.1.3.1, Supplementary Material online), we sequenced the four *O. abietinus* libraries to base coverage depths of 77×, 27×, 77×, and 44×, respectively. The amount of DNA sequences generated from each of these libraries is given in supplementary table S1, Supplementary Material online. For RNAseq data generation, poly-A mRNA was extracted from 1-μg whole-body RNA using Oligo(dT)25 Dynabeads (Life Technologies, Carlsbad, CA), followed by fragmentation of the mRNA by heat at 94 °C for 3 min (for samples with a RIN value of 3 or 3.3) or 4 min (for samples with RIN value of 6.0 and above). First-strand cDNA was synthesized using the Superscript III reverse transcriptase (Life Technologies, Carlsbad, CA) and purified using Agencourt RNAClean XP beads (Beckman Coulter, Brea, CA). During second-strand cDNA synthesis, dNTP mix containing dUTP was used to introduce strand-specificity. For Illumina paired-end library construction, the resultant cDNA was processed through end-repair and A-tailing, was ligated with Illumina PE adapters, and was then digested with 10 units of Uracil-DNA Glycosylase (New England BioLabs, Ipswich, MA). Amplification of the libraries was accomplished via 13 PCR cycles using the Phusion High-Fidelity PCR Master Mix (New England BioLabs, Ipswich, MA). We incorporated 6-bp molecular barcodes during this PCR amplification. The libraries were purified with Agencourt AMPure XP beads after each enzymatic reaction and were quality-assessed and quantified with the Agilent Bioanalyzer 2100 DNA Chip 7500 (Santa Clara, CA). The libraries were pooled in equimolar amounts prior to their sequencing. All libraries were sequenced with 101-bp read lengths on an Illumina HiSeq2000 sequencing platform. We collected the following number of reads from the whole-body RNA extract of *At. rosae*: 24,374,007 (adult male sample 1), 23,012,651 (adult male sample 2), 17,739,404 (adult female sample 1), and 8,869,760 (adult female sample 2). We collected the following number of reads from the whole-body RNA extract of *O. abietinus*: 32,320,562 (adult male) and 30,138,682 (adult female). Library preparation of the antennal transcriptomes, including poly-A enrichment, was performed using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA) according to manufacturer's instructions. The libraries were sequenced on a HiSeq2500 (Illumina, San Diego, CA), which

provided 100-bp long paired-end reads. In total, we sequenced 34,652,811 and 36,072,988 reads from antennal RNA extracts of *At. rosae* males and females, respectively. We collected a total of 20,906,900 reads from antennal RNA extracts of *O. abietinus* males. Antennal transcriptome reads were processed using CLC Genomics Workbench 7 (Qiagen, Hilden, Germany), removing adapters during read import. Cleaned reads were assembled using the de novo assembly function with its default settings, retaining only contigs of >200 bp in length.

Genome Assembly

Genome sizes and individual library coverage were estimated with jellyfish (version 2.0) (Marçais and Kingsford 2011), using 17-mers of the 180-bp genome sequencing reads. Prior to assembly, we removed all adapters from the reads with the software SeqPrep (https://github.com/jstjohn/SeqPrep). The genomes were separately assembled using ALLPATHS-LG (version 35218) (Gnerre et al. 2011) and applying the program's default parameters and the haploidy option. Contigs were scaffolded and scaffold gaps were filled with the BCM tools Atlas-Link (version 1.0) and Atlas gap-fill (version 2.2). The gene space coverage of the assemblies was assessed using BUSCO (version 1.1b1, Arthropod gene set) (Simão et al. 2015) and CEGMA (version 2.4) (Parra et al. 2007).

Automated Protein-Coding Gene Annotation

Protein-coding genes were annotated using the Maker 2.0 annotation pipeline (Cantarel et al. 2007), tuned specifically for annotating the genomes of arthropods. Specifically, the genome assembly was first subjected to de novo repeat prediction and CEGMA gene space coverage analysis; the latter for generating gene models for initial training of the ab initio gene predictors. Three rounds of training of the gene prediction programs Augustus (version 2.5.5) (Stanke et al. 2008) and SNAP (version 1.0b6) (Korf 2004) within Maker were used to infer a high-quality training set with a bootstrap method. Input protein data included 1 million peptides from a nonredundant reduction (if proteins shared > 90% amino acid sequence identity, only the first in the protein list was retained) of all Uniprot Ecdysozoa entries (1.25 million peptides; accessed July 2013), supplemented with proteomes from 18 additional species (i.e., *Strigamia maritime*, Chipman et al. 2014; *Tetranychus urticae*, Grbić et al. 2011; *Caenorhabditis elegans*, The *C. elegans* Sequencing Consortium 1998; *Loa loa*, Desjardins et al. 2013; *Trichoplax adhaerens*, Srivastava et al. 2008; *Amphimedon queenslandica*, Srivastava et al. 2010; *Strongylocentrotus purpuratus*, Sodergren et al. 2006; *Nematostella vectensis*, Putnam et al. 2007; *Branchiostoma floridae*, Putnam et al. 2008; *Ciona intestinalis*, Dehal 2002; *Ciona savignyi*, Small et al. 2007; *Homo sapiens*, International Human Genome Sequencing Consortium et al. 2001; *Mus musculus*, Mouse Genome Sequencing Consortium et al. 2002; *Capitella teleta*, Simakov et al. 2013; *Helobdella robusta*, Simakov et al. 2013; *Crassostrea gigas*, Zhang et al. 2012; *Lottia gigantean*, Simakov et al. 2013; *Schistosoma mansoni*, Berriman et al. 2009) leading to a final nonredundant peptide evidence set of 1.03 million peptides. We additionally provided MAKER RNAseq transcription data derived from two males and two females (*At. rosae*) and a single male and single female (*O. abietinus*) to identify exon–intron boundaries. We also ran a heuristic script (included as supplementary file S42, Supplementary Material online) to identify and split erroneously joined gene models.

Manual Annotations

Gene models were manually annotated with the aid of Web Apollo (Lee et al. 2013) and the i5k interface (Poelchau et al. 2015). The manual annotation process was guided by multiple intrinsic and extrinsic evidence tracks: 1) cleaned RNAseq raw read data mapped onto the genome assembly using TopHat2 (version 2.0.12) (Kim et al. 2013) with its default settings; 2) transcripts of the respective species assembled with Cufflinks (version 2.2.1) (Trapnell et al. 2010); 3) transcripts of the respective species assembled with Trinity (version trinityrnaseq_r20140413p1) (Grabherr et al. 2011) and mapped onto the genome assembly using the Exonerate (version 2.20) (Slater and Birney 2005) fork Exonerate-gff3 (https://github.com/hotdogee/exonerate-gff3) with the est2genome model (selected options were: –model est2genome –showtargetgff yes –gff3 yes–showalignment no –showvulgar no –geneseed 250 –bestn 2 –percent 50 –minintron 20 –maxintron 10,000) and marking transcripts mapping to two locations with a custom Perl script. All manually edited gene models were submitted to an automated quality control and visual inspection before being merged with the MAKER annotations into the official gene sets (OGS). The automated QC procedure (supplementary section II.3.2, Supplementary Material online) detects ~50 types of formatting errors caused by manual curation. Some errors are automatically fixed, whereas other error types need to be manually reviewed by curators or administrators. Curators were provided with a list of errors to correct in Web Apollo. After a correction period, QC reports were regenerated and the procedure repeated until no errors remained. An in-depth description of the QC procedure is available on github (https://github.com/NAL-i5K/I5KNAL_OGS/wiki).

Taxon Sampling

The genomes of *At. rosae* and *O. abietinus* were compared with those of publicly available apocritan Hymenoptera and non-Hymenoptera insects. The selected Hymenoptera comprise the honeybee, *Ap. mellifera* (Weinstock et al. 2006), the bumble bee *Bombus terrestris* (Sadd et al. 2015), the alfalfa leafcutter bee, *Megachile rotundata* (Kapheim et al. 2015),

the white-footed sweat bee, *Lasioglossum albipes* (Kocher et al. 2013), the solitary bee *Dufourea novaeangliae* (Kapheim et al. 2015), the leafcutter ant *Ac. echinatior* (Nygaard et al. 2011), the jumping ant *Harpegnathos saltator* (Bonasio et al. 2010), the Florida carpenter ant, *Camponotus floridanus* (Bonasio et al. 2010), the European paper wasp, *P. dominula* (Standage et al. 2016), and the parasitoid wasp *N. vitripennis* (Werren et al. 2010). The sampling covers the most diverged lineages and a significant fraction of the ecological width of the order. A comprehensive list of all genome assemblies and gene sets analyzed, including references, version numbers, and direct links to the data are given in supplementary table S1, Supplementary Material online.

TE Annotation

Species-specific repeat libraries were generated using RepeatModeler (version open-1.0.8) (Smit and Hubley 2015) with the program's default settings. The identified TEs were classified using a reference-based similarity search against RepBase (version update 20140131) (Jurka et al. 2005). Identified TEs were verified and annotation artifacts were removed by querying the identified sequences against the NCBI nr database (downloaded February 4, 2017) with BlastX of the BLAST+ (version 2.6.0) software suite (Camacho et al. 2009) using the software's default settings, discarding candidates without hits against known TE proteins and domains. The filtered library was finally combined with the TE sequences of RepBase (version 20140131) referring to Metazoa and used to annotate TEs with RepeatMasker (version open-4.0.5) (Smit et al. 2015) applying the software's default settings. Genomic TE coverage was calculated using the software "One code to find them all" (Bailly-Bechet et al. 2014) and intrafamily Kimura distances, used as a proxy for TE age distribution, were calculated using scripts available from the RepeatMasker (version open-4.0.5) software package. The full TE annotation pipeline was implemented in a custom shell script that is available on GitHub (github.com/mptrsen/mobilome). Testing for a correlation between genome size and TE content was done by applying a linear regression, Spearman rank sum method, and Kendall's Tau within R (R Core Team 2017). We also applied the phylogenetic independent contrast (PIC) method (Felsenstein 1985) as implemented in the ape package (Paradis et al. 2004) to control for a potential phylogenetic effect.

Comparative Analysis of Gene Structure

The structural properties of the MAKER-inferred protein-coding gene set of the two sawflies were compared with those of the selected apocritan Hymenoptera and the red flour beetle *Tribolium castaneum* (Richards et al. 2008) using COGNATE (version 1.01) (Wilbrandt et al. 2017) with the software's default settings. The *N. vitripennis* assembly version 2.1 was used instead of version 1.0 and the NCBI release 102 annotations of the *N. vitripennis* and *B. terrestris* genomes were used instead of the eviogene and Gnomon 1.0 annotations, respectively.

Orthology Prediction and Microsynteny

The predicted sawfly genes were clustered along with those of other Hymenoptera in OrthoDB (version 9.1) (Zdobnov et al. 2017) and orthology assessed at the systematic level Holometabola, with *T. castaneum* as outgroup. To investigate Hymenoptera genome evolution on a microsyntenic level, we utilized the identified SCOs and the recently published Hymenoptera divergence estimates (Peters et al. 2017). SCOs represent conserved genes that likely evolve under similar constraints (Ciccarelli 2005) and have consequently been exploited as markers to quantify genome shuffling in insects (Zdobnov and Bork 2007). Using a custom Perl script (included as supplementary file S39, Supplementary Material online), the conservation of microsynteny was inferred as the fraction of shared SCOs that retain the same neighboring SCO between two species relative to their divergences time (supplementary section II.4.5, Supplementary Material online). Positional information of the SCO was extracted from the respective OGS. GO terms were assigned to all groups of SCOs (SCOG) using the Argot2.5 web server (Lavezzo et al. 2016; http://www.medcomp.medicina.unipd.it/Argot2-5/, last accessed February 6, 2019) with the default settings, retaining only GO terms with a score of 200 or more, and InterPro2GO (Mitchell et al. 2019) (https://www.ebi.ac.uk/GOA/, last accessed February 6, 2019), using InterProScan with the default settings (version 5.33.72) (Mitchell et al. 2019). GO terms were assigned to each SCOG when shared by ten or more species in the group. Testing for GO term enrichment in the SCOGs which remained in synteny across all pairwise comparisons (754) against the background of all SCOGs (3,983) was performed using topGO's weighted Fisher test (weight01) (R package version 2.30.1) (Alexa and Rahnenfuhrer 2016) and goStats hypergeometric test (R package version 2.30.1) (Falcon and Gentleman 2007).

Gene Family and Domain Evolution

Gene family and domain evolution was analyzed with CAFE (version 4.1) (Han et al. 2013), with coupled birth and death rates, using the orthology predictions (see above) and an ultrametric tree derived from a recently published Hymenoptera phylogeny (Peters et al. 2017) as input. Following the suggestions of the authors of CAFE, the birth and death rate was determined considering only gene families with fewer than 100 copies in each species before reanalyzing the full data set with the inferred rate. Protein domains were annotated in a subset of the selected genomes with *Pfam* (version 29) (Finn et al. 2016), using the provided "pfam_scan.pl" script with the default settings. The number of unique domains and domain arrangements (the linear

sequence of domains present in a protein without repeats) occurring in each species were determined. Presence and absence of domains among species were inferred using a custom python script (pyDomrates; https://github.com/sklas/pyDomrates) and the *ETE3* python module (Huerta-Cepas et al. 2016). The gain and loss of domains at nodes of the tree were inferred applying the Fitch parsimony optimality criterion. Domains are considered as gained at a node if they were inferred to not have been present at the parent node. Likewise, domains are considered as lost if they were inferred to have been present only at the parent node.

Major Royal Jelly Proteins

DNA sequences of specific exons of *yellow* and *mrjp*/-like genes of *Ap. mellifera* and *N. vitripennis* were used as query to search them with the TBlastX search algorithm with the default settings (BLAST web server hosted by the NCBI) against the reference genome assemblies of *At. rosae*, *O. abietinus*, and *L. albipes*. All found coding sequences were manually curated and aligned along those of *Ap. mellifera*, *B. terrestris*, *M. rotundata*, *Du. novaeangliae*, *P. dominula*, *Ac. echinatior*, *Ca. floridanus*, *H. saltator*, *Linepithema humile* (Smith, Zimin, et al. 2011), *N. vitripennis*, *T. castaneum*, and *Zootermopsis nevadensis* (Terrapon et al. 2014) (supplementary file S13, Supplementary Material online) at the translational level with ClustalW implemented in MEGA (version 6.0.6) (Tamura et al. 2013) with the default settings. We inferred a maximum-likelihood tree from the aligned amino acid sequences, using the WAG+F+R7 amino acid substitution model. Branch support was assessed from 1,000 nonparametric bootstrap replicates. Maximum-likelihood tree reconstruction was performed in IQ-TREE (version 1.6.6) (Nguyen et al. 2015) and the best-fitting model was selected using ModelFinder (Kalyaanamoorthy et al. 2017) as implemented in IQ-TREE. Topology tests were done in IQ-TREE (version 1.6.8) using 1) likelihood-mapping (Strimmer and von Haeseler 1997) with four clusters (MRJPls, Y-e3, Y-x2, and all remaining Yellow proteins) and 2) an approximate unbiased test (Shimodaira 2002), testing the inferred ML-tree (MRJPls and Y-x2 as sister-groups) against the alternative hypothesis (MRJPls and Y-e3 as sister-groups) using 1 million RELL replicates.

Immune Genes

A set of immune genes was selected based on the Insect Innate Immunity Database (IIID) (Brucker et al. 2012) and modified according to previous studies on Hymenoptera (Evans et al. 2006; Gadau et al. 2012; Barribeau et al. 2015). Immune genes were identified with the aid of profile hidden Markov models (HMM), utilizing reference immune response-related amino acid sequences obtained from OrthoDB (Version 9) and the NCBI protein database (including RefSeq; Pruitt et al. 2012). All amino acid sequences were aligned with MAFFT with the default settings (version 7.130) (Katoh and Standley 2013) and the HMM profiles were inferred with the software HMMER (version 3.1b1) (http://hmmer.org/) with the default settings. The HMM profiles were searched against the predicted proteins with the HMM search tool hmmsearch with the default settings. All immunity gene candidates were evaluated with a PFAM sequence search (https://pfam.xfam.org) to exclude false positives, retaining only candidate sequences with hits against known immune genes.

Vision Genes

Opsin-coding genes were identified using amino acid reference sequences of the corresponding proteins in *Ap. mellifera*, *Drosophila melanogaster*, and *T. castaneum* obtained from UniProt. Reference sequences were searched against the genome assemblies with the TBlastN software of the BLAST+ software suite (version 2.6.0) with the default settings. Candidate orthologs were reciprocally searched with the aid of the BLAST+ software suite, with the default settings, against the *Ap. mellifera*, *Dr. melanogaster*, and *T. castaneum* genome assemblies to sort out false positives. Finally, all verified opsin genes were manually curated within Web Apollo. Opsin amino acid sequences of *At. rosae* and *O. abietinus* were aligned to those of *Ap. mellifera*, *Dr. melanogaster*, and *N. vitripennis* (Pultz and Leaf 2003) using ClustalW (v2.1) (Larkin et al. 2007) with the default settings. Ambiguous alignment regions were excluded using the software TrimAl (version 1.3) (Capella-Gutierrez et al. 2009), implemented on the Phylemon 2.0 server (Sanchez et al. 2011) and applying the "Automated 1" settings. A maximum-likelihood tree was estimated with the MEGA software (version 6.0) and applying the JTT+G amino acid substitution model. Branch support values were estimated from 500 nonparametric bootstrap replicates.

Metabolism

We functionally annotated all predicted proteins of *At. rosae, O. abietinus*, and *N. vitripennis* with the CycADS pipeline (version 1.32) (Vellozo et al. 2011) (supplementary section II.5.10, Supplementary Material online) with the default settings. CycADS is an annotation database system that collects functional annotations predicted by multiple computational methods including BLAST2Go (version 2.5) (Götz et al. 2008), InterProScan (version 5.0) (Mitchell et al. 2019), Kaas-Kegg server (version 2.0) (Moriya et al. 2007), and Priam (March 13. release) (Claudel-Renard et al. 2003). Predicted EC numbers and Gene Ontology terms (GO) collected by CycADS were then processed with the Pathway Tools software (Karp et al. 2016) to infer enzymatic reactions and metabolic pathways that were finally manually curated and compared.

Storage Proteins

Hexamerins of selected Hymenoptera (*Ac. echinatior*, *Ap. mellifera*, *Ca. floridanus*, *H. saltator*) and an outgroup species, the termite *Z. nevadensis*, were downloaded from UniProt and used to identify hexamerins in the *At. rosae* and *O. abietinus* genomes using the software BLAT (Kent 2002) implemented in the i5k@NAL workspace, using an e-value cut-off of 1e-10. The reference sequences were aligned against the newly identified hexamerins of the two sawflies with MAFFT (version 7) using the E-INS-i algorithm with the default settings. The multiple amino acid sequence alignment was further processed with GBlocks (version 0.91b) (Castresana 2000) with the default settings. A maximum-likelihood tree was inferred using IQ-TREE (version 1.6.6) applying the best-fitting amino acid substitution model after the BIC criterion (LG+G4) as determined by ModelFinder. Branch support values were estimated from 1,000 nonparametric bootstrap replicates.

Odorant and Gustatory Receptors

Initial candidate genes were identified by querying reference amino acid sequences from Hymenoptera (Zhou et al. 2015; Robertson et al. 2018) against the MAKER-inferred gene set and the genome assemblies using TBlastN (version 2.2.31) with the default settings. Candidate gene models were manually annotated or corrected in Web Apollo considering raw reads and assembled transcripts of the antennal transcriptomes which were mapped against the genomes of the respective species using the "map to reference" function in CLC Genomics Workbench 7 (Qiagen, Hilden, Germany) with the program's default settings. Annotated gene models were queried against the assemblies along with those of other Hymenoptera to identify additional genes potentially missed by the initial annotation. Candidate nucleotide sequences were subsequently searched against the NCBI nr database with TBlastX to eliminate false positives with the default settings. Predicted amino acid sequences were aligned to those of *Ac. echinatior*, *Ap. mellifera*, *N. vitripennis* (Zhou et al. 2015), and *Ce. cinctus* (Robertson et al. 2018) using MUSCLE (version 3.8.31) (Edgar 2004) with the default settings. All resulting alignments were visually inspected and, if necessary, manually curated. Maximum-likelihood phylogenies were built using PhyML (version 3.0) (Guindon et al. 2010) under the best-fitting substitution model as determined by SMS (Lefort et al. 2017). Branch support was estimated through an approximate likelihood-ratio test (Anisimova and Gascuel 2006). All phylogenetic trees were visualized with FigTree (version 1.4.2) (http://tree.bio.ed.ac.uk/software/figtree/).

Software Availability

The custom Perl script used to infer pairwise microsynteny is provided in the Supplementary Material online (supplementary file S39, Supplementary Material online).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This genome sequencing project contributes to the i5K initiative, whose target is to sequence the genomes of 5,000 arthropods (Robinson et al. 2011; Evans et al. 2013). We thank the staff at the Baylor College of Medicine Human Genome Sequencing Center for their contributions. E.B.B., Al.D., S.K., C.M., B.M., O.N., J.P.O., R.S.P., Ma.P., L.P., J.W., and T.Z. acknowledge the Leibniz Graduate School on Genomic Biodiversity research. B.M.v.R. thanks the Fraunhofer IME (Gießen) for infrastructure. B.M. and O.N. thank Claudia Etzbauer and Sandra Kukowka (ZFMK, Bonn) for technical support. O.N. and J.P.O. thank Caroline Müller (University of Bielefeld) for providing samples of *At. rosae* to take photographs of this species. O.N. furthermore thanks Stefan Tischendorf (Darmstadt) for sending detailed information where and when to collect *O. abietinus* in the field. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer. Genome sequencing, assembly, and annotation were funded by a Grant No. U54 HG003273 from the National Human Genome Research Institute (NHGRI to R.A.G.). B.M., V.C.M., O.N., Ma.P., B.M.v.R., and J.W. additionally acknowledge the German Research Foundation (DFG) for financial support (MI 649/x-1; NI 1387/2-1; NI 1387/3-1; RE 3454/2-1; RE 3454/6-1). J.H.W. thanks the US National Science Foundation (DEB 1257053 and IOS 1456233). R.M.W. acknowledges support from the Swiss National Science Foundation (PP00P3_170664). V.C.M. and O.N. acknowledge the support of the Freiburg Galaxy Team and Prof. Rolf Backofen, Bioinformatics, University of Freiburg, Germany, funded by Collaborative Research Centre 992 Medical Epigenetics (DFG Grant No. SFB 992/1 2012) and German Federal Ministry of Education and Research (BMBF Grant No. 031 A538A RBC [de.NBI]). This project was also supported by Det Frie Forskningsråd (Grant No. 7014-0008B to C.J.P.G.).

Author Contributions

Conceived the project: B.M., J.P.O., M.H., O.N., R.S.P., and S.R. Lead authors: B.M., J.P.O., and O.N. Major contributors:

B.M., B.O., J.P.O., M.H., O.N., and S.R. Project coordinator: J.P.O. Project management: J.P.O., M.F.P., R.A.G., S.D., and S.R. Antioxidants: E.C.J. and J.B.B. Aquaporins: A.L.S. and J.B.B. Automated annotation: D.H. and S.R. Autophagy genes: E.M.S. and J.B.B. BUSCO: E.Z., F.A.S., P.I., and R.M.W. CEGMA: J.P.O. and O.N. Chemoreceptor repertoire: B.H., E.G.W., E.J.J., E.P., and N.M. Cuticular proteins: A.J.R. and J.B.B. Metabolism: F.C., Hu.C., J.P.O., N.P., and P.B.P. Cysteine peptidases: A.G.M., B.O., and E.N.E. Desaturases: E.C. Dosage compensation: D.S.Y., M.H., and M.S. Developmental genes: A.G.C., E.D., J.S., M.L., O.T., and P.D. DNA methylation: B.M., P.P., and S.G. Elongases: O.N. and V.C.M. Facilitated manual gene curation: M.M.T. Gene family evolution: B.M. and J.P.O. Gene structure comparison: B.M., J.P.O., J.W., and O.N. GO-term: R.M.W. and M.R. Heat shock proteins: E.C.J., E.M.S., and J.B.B. Hexamerins: C.P. Immune genes: L.P. and P.L. Insulator proteins: L.V., O.N., and T.P. Lateral gene transfer: Am.D., J.H.W., M.H., and O.N. Major royal jelly proteins: A.B. Metabolism of host plant toxins: D.S.Y., M.H., and M.S. Metallopeptidases: L.P. and S.M. Microsynteny: B.M., J.P.O., and O.N. Mitochondrial genomes: Al.D., J.P.O., and O.N. Neuropeptides and their G protein-coupled receptors: A.G.S., C.J.P.G., E.T., and F.H. Noncoding RNA: Al.D. and T.Z. Orthology: E.Z., P.I., and R.M.W. OXPHOS genes: J.D.G. Protein domains: E.B.B. and S.K. Quality control and official gene set generation: C.C., M.C., and M.F.P. Repeats and transposable elements: C.M., Ma.P., and J.P.O. RNAi: Al.D., D.D., and O.N. Sequencing and assembly: D.H., D.M.M., H.D., Hs.C., H.V.D., J.P.O., J.Q., K.C.W., O.N., M.H., S.C.M., S.L.L., and Y.H. Sex determination: D.S.Y., E.G., L.W.B., L.v.d.Z., M.H., and M.S. Silk proteins: D.S.Y., M.H., and M.S. Variable sites: L.P. Venom proteins: B.Mv.R. Vision genes: C.G., J.W.J., and M.F. Web Apollo: C.C.

Literature Cited

Abdalsamee MK, Müller C. 2012. Effects of indole glucosinolates on performance and sequestration by the sawfly *Athalia rosae* and consequences of feeding on the plant defense system. J Chem Ecol. 38(11):1366–1375.

Abe M. 1988. A biosystematic study of the genus *Athalia* Leach of Japan (Hymenoptera: tenthredinidae). Esakia 26:91–131.

Aguiar AP, et al. 2013. Order Hymenoptera. Zootaxa 3703(1):51–62.

Ahnlund H, Ronquist F. 2002. Den röda parasitväxtstekelns (*Orussus abietinus*) biologi och förekomst i Norden. Entomol Tidskr. 122:1–10.

Alexa A, Rahnenfuhrer J. 2016. topGO: Enrichment analysis for gene ontology. R package version 2.38.1.

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol. 55(4):539–552.

Ardila-Garcia AM, Umphrey GJ, Gregory TR. 2010. An expansion of the genome size dataset for the insect order Hymenoptera, with a first test of parasitism and eusociality as possible constraints. Insect Mol Biol. 19(3):337–346.

Bailly-Bechet M, Haudry A, Lerat E. 2014. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. Mob DNA. 5(1):13.

Bao R, Friedrich M. 2009. Molecular evolution of the *Drosophila* retinome: exceptional gene gain in the higher Diptera. Mol Biol Evol. 26(6):1273–1287.

Barribeau SM, et al. 2015. A depauperate immune repertoire precedes evolution of sociality in bees. Genome Biol. 16(1):83.

Beres BL, Dosdall LM, Weaver DK, Cárcamo HA, Spaner DM. 2011. Biology and integrated management of wheat stem sawfly and the need for continuing research. Can Entomol. 143(2):105–125.

Berriman M, et al. 2009. The genome of the blood fluke *Schistosoma mansoni*. Nature 460(7253):352–358.

Bonasio R, et al. 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. Science 329(5995):1068–1071.

Branstetter MG, et al. 2018. Genomes of the Hymenoptera. Curr Opin Insect Sci. 25:65–75.

Brucker RM, Funkhouser LJ, Setia S, Pauly R, Bordenstein SR. 2012. Insect Innate Immunity Database (IIID): an annotation tool for identifying immune genes in insect genomes. PLoS One 7(9):e45125.

Buttstedt A, Moritz RFA, Erler S. 2014. Origin and function of the major royal jelly proteins of the honeybee (*Apis mellifera*) as members of the *yellow* gene family. Biol Rev. 89(2):255–269.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1):421.

Cantarel BL, et al. 2007. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18(1):188–196.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17(4):540–552.

Chipman AD, et al. 2014. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. PLoS Biol. 12(11):e1002005.

Ciccarelli FD. 2005. Complex genomic rearrangements lead to novel primate gene function. Genome Res. 15(3):343–351.

Claudel-Renard C, Chevalet C, Faraut T, Kahn D. 2003. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res. 31(22):6633–6639.

Cristino AS, et al. 2010. Organization, evolution and transcriptional profile of hexamerin genes of the parasitic wasp *Nasonia vitripennis* (Hymenoptera: Pteromalidae). Insect Mol Biol. 19:137–146.

Dehal P. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science 298(5601):2157–2167.

Desjardins CA, et al. 2013. Genomics of *Loa loa*, a *Wolbachia*-free filarial parasite of humans. Nat Genet. 45(5):495–501.

Desjardins CA, Perfectti F, Bartos JD, Enders LS, Werren JH. 2010. The genetic basis of interspecies host preference differences in the model parasitoid *Nasonia*. Heredity 104(3):270–277.

Drapeau MD, Albert S, Kucharski R, Prusko C, Maleszka R. 2006. Evolution of the Yellow/Major Royal Jelly protein family and the emergence of social behavior in honey bees. Genome Res. 16(11):1385–1394.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Engström PG, Sui SJH, Drivenes Ø, Becker TS, Lenhard B. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res. 17(12):1898–1908.

Evans JD, Brown SJ, Hackett KJJ, Robinson G, Richards S. 2013. The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered. 104:595–600.

Evans JD, et al. 2006. Immune pathways and defence mechanisms in honey bees *Apis mellifera*. Insect Mol Biol. 15(5):645–656.

Falcon S, Gentleman R. 2007. Using GOstats to test gene lists for GO term association. Bioinformatics 23(2):257–258.

Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat. 125(1):1–15.
Finn RD, et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44(D1):D279–D285.
Gadau J, et al. 2012. The genomic impact of 100 million years of social evolution in seven ant species. Trends Genet. 28(1):14–21.
Gnerre S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 108(4):1513–1518.
Götz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36(10):3420–3435.
Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7):644–652.
Grbić M, et al. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature 479(7374):487–492.
Gregory TR. 2018. Animal genome size database. Available from: http://www.genomesize.com. Accessed December 2019.
Grimaldi DA, Engel MS. 2005. Evolution of the insects. Cambridge: Cambridge University Press.
Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59(3):307–321.
Hagner-Holler S, Pick C, Girgenrath S, Marden JH, Burmester T. 2007. Diversity of stonefly hexamerins and implication for the evolution of insect storage proteins. Insect Biochem Mol Biol. 37(10):1064–1074.
Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. 30(8):1987–1997.
Hanes J, Šimuth J. 1992. Identification and partial characterization of the major royal jelly protein of the honey bee (*Apis mellifera* L.). J Apic Res. 31(1):22–26.
Hanrahan SJ, Johnston JS. 2011. New genome size estimates of 134 species of arthropods. Chromosome Res. 19(6):809–823.
Henze MJ, Oakley TH. 2015. The dynamic evolutionary history of pancrustacean eyes and opsins. Integr Comp Biol. 55(5):830–842.
Hu X, Leming MT, Whaley MA, O'Tousa JE. 2014. Rhodopsin coexpression in UV photoreceptors of *Aedes aegypti* and *Anopheles gambiae* mosquitoes. J Exp Biol. 217(6):1003–1008.
Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 33(6):1635–1638.
International Human Genome Sequencing Consortium, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
Jervis MA, Ellers J, Harvey JA. 2008. Resource acquisition, allocation, and utilization in parasitoid reproductive strategies. Annu Rev Entomol. 53(1):361–385.
Jervis MA, Kidd NAC, Fitton MG, Huddleston T, Dawah HA. 1993. Flower-visiting by hymenopteran parasitoids. J Nat Hist. 27(1):67–105.
Jones WD, Cayirlioglu P, Kadow IG, Vosshall LB. 2007. Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. Nature 445(7123):86–90.
Jurka J, et al. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110(1–4):462–467.
Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 14(6):587–589.
Kapheim KM, et al. 2015. Genomic signatures of evolutionary transitions from solitary to group living. Science 348(6239):1139–1143.
Karp PD, et al. 2016. Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. Brief Bioinform. 17(5):877–890.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.
Kawecki TJ. 1998. Red queen meets Santa Rosalia: arms races and the evolution of host specialization in organisms with parasitic lifestyles. Am Nat. 152(4):635–651.
Kent CF, Minaei S, Harpur BA, Zayed A. 2012. Recombination is associated with the evolution of genome structure and worker behavior in honey bees. Proc Natl Acad Sci U S A. 109(44):18012–18017.
Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12(4):656–664.
Kim D, et al. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14(4):R36.
Kim S, et al. 2016. Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly. Genome Biol. 17(1):211.
Kocher SD, et al. 2013. The draft genome of a socially polymorphic halictid bee, *Lasioglossum albipes*. Genome Biol. 14(12):R142.
Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5(1):59.
Kupke J, Spaethe J, Mueller MJ, Rössler W, Albert Š. 2012. Molecular and biochemical characterization of the major royal jelly protein in bumblebees suggest a non-nutritive function. Insect Biochem Mol Biol. 42(9):647–654.
Kwon JY, Dahanukar A, Weiss LA, Carlson JR. 2007. The molecular basis of CO_2 reception in *Drosophila*. Proc Natl Acad Sci U S A. 104(9):3574–3578.
Larkin MA, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947–2948.
Lavezzo E, Falda M, Fontana P, Bianco L, Toppo S. 2016. Enhancing protein function prediction with taxonomic constraints—the Argot2.5 web server. Methods 93:15–23.
Lee E, et al. 2013. Web Apollo: a web-based genomic annotation editing platform. Genome Biol. 14(8):R93.
Lees JG, Dawson NL, Sillitoe I, Orengo CA. 2016. Functional innovation from changes in protein domains and their combinations. Curr Opin Struct Biol. 38:44–52.
Lefort V, Longueville J-E, Gascuel O. 2017. SMS: smart model selection in PhyML. Mol Biol Evol. 34(9):2422–2424.
Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27(6):764–770.
Martins JR, Nunes FMF, Cristino AS, Simões ZLP, Bitondi M. 2010. The four hexamerin genes in the honey bee: structure, molecular evolution and function deduced from expression patterns in queens, workers and drones. BMC Mol Biol. 11(1):23.
Mattson WJ. 1980. Herbivory in relation to plant nitrogen content. Annu Rev Ecol Syst. 11(1):119–161.
Mine S, Sumitani M, Aoki F, Hatakeyama M, Suzuki MG. 2017. Identification and functional characterization of the sex-determining gene doublesex in the sawfly, *Athalia rosae* (Hymenoptera: Tenthredinidae). Appl Entomol Zool. 52(3):497–509.
Mitchell AL, et al. 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 47(D1):D351–D360.
Moore AD, Bornberg-Bauer E. 2012. The dynamics and evolutionary potential of domain loss and emergence. Mol Biol Evol. 29(2):787–796.
Moreno M, Giralt E. 2015. Three valuable peptides from bee and wasp venoms for therapeutic and biotechnological use: melittin, apamin and mastoparan. Toxins 7(4):1126–1150.
Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35(Web Server):W182–W185.
Mouse Genome Sequencing Consortium, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.

Mrinalini, Werren JH. 2017. Parasitoid wasps and their venoms. In: Gopalakrishnakone P, Malhotra A, editors. Evolution of venomous animals and their toxins. Dordrecht: Springer. p. 187–212.
Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32(1):268–274.
Ni JD, Baik LS, Holmes TC, Montell C. 2017. A rhodopsin in the brain functions in circadian photoentrainment in *Drosophila*. Nature 545(7654):340–344.
Niehuis O, et al. 2010. Recombination and its impact on the genome of the haplodiploid parasitoid wasp *Nasonia*. PLoS One 5(1):e8597.
Nygaard S, et al. 2011. The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. Genome Res. 21(8):1339–1348.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20(2):289–290.
Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23(9):1061–1067.
Pask GM, et al. 2017. Specialized odorant receptors in social insects that detect cuticular hydrocarbon cues and candidate pheromones. Nat Commun. 8(1):297.
Peitsch D, et al. 1992. The spectral input systems of hymenopteran insects and their receptor-based colour vision. J Comp Physiol A. 170(1):23–40.
Peters RS. 2010. Host range and offspring quantities in natural populations of *Nasonia vitripennis* (Walker, 1836) (Hymenoptera: Chalcidoidea: Pteromalidae). J Hymenopt Res. 19:128–138.
Peters RS, et al. 2014. The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. BMC Evol Biol. 14(1):52.
Peters RS, et al. 2017. Evolutionary history of the Hymenoptera. Curr Biol. 27(7):1013–1018.
Petersen M, et al. 2019. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC Evol Biol. 19(1):11.
Poelchau M, et al. 2015. The i5k Workspace@NAL-enabling genomic data access, visualization and curation of arthropod genomes. Nucleic Acids Res. 43(D1):D714–D719.
Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 40(D1):D130–D135.
Pultz MA, Leaf DS. 2003. The jewel wasp *Nasonia*: querying the genome with haplo-diploid genetics. Genesis 35(3):185–191.
Putnam NH, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317(5834):86–94.
Putnam NH, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453(7198):1064–1071.
Quicke D. 1997. Parasitic wasps. London: Chapman & Hall.
R Core Team. 2017. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from: https://www.R-project.org/.
Rappoport N, Linial M. 2015. Trends in genome dynamics among major orders of insects revealed through variations in protein families. BMC Genomics 16(1):583.
Richards S, et al. 2008. The genome of the model beetle and pest *Tribolium castaneum*. Nature 452(7190):949–955.
Robertson HM, et al. 2018. Genome sequence of the wheat stem sawfly, *Cephus cinctus*, representing an early-branching lineage of the Hymenoptera, illuminates evolution of hymenopteran chemoreceptors. Genome Biol Evol. 10(11):2997–3011.
Robertson HM, Gadau J, Wanner KW. 2010. The insect chemoreceptor superfamily of the parasitoid jewel wasp *Nasonia vitripennis*. Insect Mol Biol. 19:121–136.
Robertson HM, Wanner KW. 2006. The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. Genome Res. 16(11):1395–1403.
Robinson GE, et al. 2011. Creating a buzz about insect genomes. Science 331(6023):1386–1386.
Sackton TB, Werren JH, Clark AG. 2013. Characterizing the infection-induced transcriptome of *Nasonia vitripennis* reveals a preponderance of taxonomically-restricted immune genes. PLoS One 8(12):e83984.
Sadd BM, et al. 2015. The genomes of two key bumblebee species with primitive eusocial organization. Genome Biol. 16(1):76.
Sanchez R, et al. 2011. Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. Nucleic Acids Res. 39(Suppl):W470–W474.
Sáringer G. 1974. Problems with *Athalia rosae* L. (Hym., Tenthredinidae) in Hungary. In: Proceed 4. Giessen: Intern Rapskongress. p. 575–578.
Savard J, et al. 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of holometabolous insects. Genome Res. 16(11):1334–1338.
Schmid-Hempel P. 2011. Evolutionary parasitology: the integrated study of infections, immunology, ecology, and genetics. Oxford: Oxford University Press.
Sekine K, Furusawa T, Hatakeyama M. 2015. The *boule* gene is essential for spermatogenesis of haploid insect male. Dev Biol. 399(1):154–163.
Senthilan PR, Helfrich-Förster C. 2016. Rhodopsin 7–the unusual rhodopsin in *Drosophila*. PeerJ 4:e2427.
Sharkey MJ. 2007. Phylogeny and classification of Hymenoptera. Zootaxa 1668(1):521–548.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51(3):492–508.
Simakov O, et al. 2013. Insights into bilaterian evolution from three spiralian genomes. Nature 493(7433):526–531.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.
Slansky FJ. 1986. Nutritional ecology of endoparasitic insects and their hosts: an overview. J Insect Physiol. 32(4):255–261.
Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6(1):31.
Small KS, Brudno M, Hill MM, Sidow A. 2007. A haplome alignment and reference sequence of the highly polymorphic *Ciona savignyi* genome. Genome Biol. 8(3):R41.
Smit A, Hubley R. 2015. RepeatModeler Open-1.0. Available from: http://wwww.repeatmasker.org. Accessed December 18, 2016.
Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. Available from: http://www.repeatmasker.org. Accessed December 18, 2016.
SmithCD, Zimin A, et al. 2011. Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). Proc Natl Acad Sci U S A. 108(14):5673–5678.
Smith CR, SmithCD, et al. 2011. Draft genome of the red harvester ant *Pogonomyrmex barbatus*. Proc Natl Acad Sci U S A. 108(14):5667–5672.
Snodgrass RE. 1925. Anatomy and physiology of the honeybee. New York: Mcgeaw-Hallbook Company.
Sodergren E, et al. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. Science 314(5801):941–952.
Srivastava M, et al. 2008. The *Trichoplax* genome and the nature of placozoans. Nature 454(7207):955–960.
Srivastava M, et al. 2010. The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature 466(7307):720–726.
Standage DS, et al. 2016. Genome, transcriptome and methylome sequencing of a primitively eusocial wasp reveal a greatly reduced DNA methylation system in a social insect. Mol Ecol. 25(8):1769–1784.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24(5):637–644.
Steidle JLM, Schöller M. 1997. Olfactory host location and learning in the granary weevil parasitoid *Lariophagus distinguendus* (Hymenoptera: Pteromalidae). J Insect Behav. 10(3):331–342.
Strimmer K, von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci U S A. 94(13):6815–6819.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 30(12):2725–2729.
Tavares MG, Carvalho CR, Soares F. 2010. Genome size variation in *Melipona* species (Hymenoptera: Apidae) and sub-grouping by their DNA content. Apidologie 41(6):636–642.
Terrapon N, et al. 2014. Molecular traces of alternative social organization in a termite genome. Nat Commun. 5(1):3636.
The C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282:2012–2018.
Trapnell C, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28(5):511–515.
Velarde RA, Sauer CD, Walden KKO, Fahrbach SE, Robertson HM. 2005. Pteropsin: a vertebrate-like non-visual opsin expressed in the honey bee brain. Insect Biochem Mol Biol. 35(12):1367–1377.
Vellozo AF, et al. 2011. CycADS: an annotation database system to ease the development and update of BioCyc databases. Database 2011(0):bar008–bar008.
Vilhelmsen L. 2003. Larval anatomy of Orussidae (Hymenoptera). J Hymenopt Res. 12:346–354.
Vilhelmsen L, Isidoro N, Romani R, Basibuyuk HH, Quicke DLJ. 2001. Host location and oviposition in a basal group of parasitic wasps: the subgenual organ, ovipositor apparatus and associated structures in the Orussidae (Hymenoptera, Insecta). Zoomorphology 121(2):63–84.
Vilhelmsen L, Turrisi GF. 2011. Per arborem ad astra: morphological adaptations to exploiting the woody habitat in the early evolution of Hymenoptera. Arthropod Struct Dev. 40(1):2–20.
Vinson SB, Iwantsch GF. 1980. Host regulation by insect parasitoids. Q Rev Biol. 55(2):143–165.
Visser B, et al. 2010. Loss of lipid synthesis as an evolutionary consequence of a parasitic lifestyle. Proc Natl Acad Sci U S A. 107(19):8677–8682.
Wakakuwa M, Kurasawa M, Giurfa M, Arikawa K. 2005. Spectral heterogeneity of honeybee ommatidia. Naturwissenschaften 92(10):464–467.
Weihrauch D, Donini A, O'Donnell MJ. 2012. Ammonia transport by terrestrial and aquatic insects. J Insect Physiol. 58(4):473–487.
Weinstock GM, et al. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443:931–949.
Werren JH, et al. 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science 327(5963):343–348.
Whitfield JB. 1998. Phylogeny and evolution of host-parasitoid interactions in Hymenoptera. Annu Rev Entomol. 43(1):129–151.
Wilbrandt J, Misof B, Niehuis O. 2017. COGNATE: comparative gene annotation characterizer. BMC Genomics 18(1):535.
Wilfert L, Gadau J, Schmid-Hempel P. 2007. Variation in genomic recombination rates among animal taxa and the case of social insects. Heredity 98(4):189–197.
Yamamoto DS, Sumitani M, Tojo K, Lee JM, Hatakeyama M. 2004. Cloning of a decapentaplegic orthologue from the sawfly, *Athalia rosae* (Hymenoptera), and its expression in the embryonic appendages. Dev Genes Evol. 214(3):128–133.
Zdobnov EM, Bork P. 2007. Quantification of insect genome divergence. Trends Genet. 23(1):16–20.
Zdobnov EM, et al. 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res. 45(D1):D744–D749.
Zhang G, et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490(7418):49–54.
Zhou X, et al. 2012. Phylogenetic and transcriptomic analysis of chemosensory receptors in a pair of divergent ant species reveals sex-specific signatures of odor coding. PLoS Genet. 8(8):e1002930.
Zhou X, et al. 2015. Chemoreceptor evolution in Hymenoptera and its implications for the evolution of eusociality. Genome Biol Evol. 7(8):2407–2416.

Associate editor: Dennis Lavrov